AD_____

Award Number: DAMD17-98-1-8003

TITLE: New Strategies for Drug Discovery and Development for *Plasmodium falciparum*

PRINCIPAL INVESTIGATOR: Dyann F. Wirth, PhD

CONTRACTING ORGANIZATION: Harvard University
Cambridge, Massachusetts 02138

REPORT DATE: January 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2001	Final (16 Dec 97 - 15 Dec 00)

4. TITLE AND SUBTITLE
New Strategies for Drug Discovery and Development for *Plasmodium falciparum*

5. FUNDING NUMBERS
DAMD17-98-1-8003

6. AUTHOR(S)
Dyann F. Wirth, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Harvard University
Cambridge, Massachusetts 02138
E-Mail: dfwirth@hsph.harvard.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)*** Malaria continues as a major health threat throughout the tropical world and potential demand for antimalarials is higher than for any other medication yet the world faces a crisis-drug resistance is emerging and spreading faster than drugs are being developed and the flow in the pipeline of new drugs has all but stopped. This represents a particular threat to the US Military. In a short time there may be parts of the world where no effective antimalarial drug is available. The recent emergence of multidrug resistant malaria parasites has intensified this problem. Recognizing this emerging crisis, it is necessary to identify new strategies for the identification and development of new antimalarials. The goal of this work is the development of a framework for antimalarial drug development into the 21st century. A new strategy for drug development is urgently needed. Current drugs are based on a small number of target molecules or lead compounds and in most cases the target of drug action is yet to be identified. Resistance is emerging rapidly and the mechanisms of resistance are poorly understood. The identification of new targets or new candidate drugs based on an understanding of the parasite biology are key elements in this new strategy. Clearly the development of a new antimalarial will require both basic and applied research working in concert with one another. The goal of this work is to use a molecular genetic approach both in the identification of new drug targets and in the investigation of mechanisms of drug resistance. The research has focused on the two objectives, namely the analysis of critical genes in the *Plasmodium falciparum* for their role in drug resistance and as potential new drug targets, including the analysis of gene expression in response to drug treatment using the method of Serial Analysis of Gene Expression and the use of DNA Chip technology in the analysis of the yeast heterologous system. These approaches complement ongoing work and will provide us with new insights into drug resistance and provide excellent tools for the identification of potential new drug targets.

14. SUBJECT TERMS Drug resistance, malaria, transformation, drug chemotherapy, multidrug resistance (mdr), yeast expression, serial analysis of gene expression (SAGE), yeast microarray

15. NUMBER OF PAGES 120

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

Table of Contents

(4) Introduction

Malaria continues as a major health threat throughout the tropical world and potential demand for antimalarials is higher than for any other medication yet the world faces a crisis-drug resistance is emerging and spreading faster than drugs are being developed and the flow in the pipeline of new drugs has all but stopped. This represents a particular threat to the US Military. In a short time there may be parts of the world where no effective antimalarial drug is available. The recent emergence of multidrug resistant malaria parasites has intensified this problem. Recognizing this emerging crisis, it is necessary to identify new strategies for the identification and development of new antimalarials. The goal of this work is the development of a framework for antimalarial drug development into the 21st century.

A new strategy for drug development is urgently needed. Current drugs are based on a small number of target molecules or lead compounds and in most cases the target of drug action is yet to be identified. Resistance is emerging rapidly and the mechanisms of resistance are poorly understood. The identification of new targets or new candidate drugs based on an understanding of the parasite biology are key elements in this new strategy. Clearly the development of a new antimalarial will require both basic and applied research working in concert with one another.

The goal of this work is to use a molecular genetic approach both in the identification of new drug targets and in the investigation of mechanisms of drug resistance. There are two parallel approaches being developed, one the development and characterization of a homologous transformation system and two the development of a heterologous expressions system in yeast for potential drug target enzymes. The yeast expression system should allow rapid screening of new drugs, greatly increasing the rate at which new antimalarials can be tested and developed. Both of these approaches are based on the functional analysis of malaria genes with goal of using this information in the identification and development of new antimalarial drugs. The development of these tools should facilitate future drug development and allow us to translate our molecular genetic knowledge into the practical identification and development of new antimalarials. This is a new strategy and it is being applied because of the crisis facing us in antimalarial drugs. The previous strategy, namely lead directed screening must be supplemented by new strategies or we will be faced with multiresistant *Plasmodium falciparum* and no drugs to treat it.

Malaria represents a major and increasing threat to the U.S. Military. Many of the sites of current or potential U.S. Military involvement are endemic for malaria and in several sites, multidrug resistant *P. falciparum* represents a major problem especially for non-immune military personnel. Current drugs available to the U.S. Military are quickly losing their effectiveness because of emerging and spreading drug resistance. This work is directed both at identifying new drugs and drug targets, but equally importantly toward an understanding of drug resistance mechanisms with the goal of preventing or overcoming drug resistance in the malaria parasite.

(5) Body

During the grant period, the research has focused on the two objectives, namely the analysis of critical genes in the Plasmodium falciparum for their role in drug resistance and as potential new drug targets using both the homologous *P. falciparum* system and the heterologous yeast system. We have initiated experiments during this grant period which take a alternate technical approach to achieve the goals in our statement of work and represent applications of new technology which did not exist at the time of our original planning process. These include the analysis of gene expression in response to drug treatment using the method of Serial Analysis of Gene Expression and the use of DNA Chip technology in the analysis of the yeast heterologous system. These approaches complement ongoing work and will provide us with new insights into drug resistance and provide excellent tools for the identification of potential new drug targets. Summaries of the ongoing work, including recent data are included for each of the projects. This report is for the entire grant period and includes information from previous annual and other interim reports.

5.1 Functional analysis of putative drug resistance genes and new drug target genes in the malaria parasite through the further development of a transformation system for the malaria parasite including:

1. Development of methods to express and modify parasite genes
2. Development of methods for targeted gene disruption

The overall goal of this work is to understand gene expression the parasite, in particular, the expression of genes critical for drug response and resistance. This work will also lead to the development of methods to identify critical genes as future drug targets. One of the key obstacles hindering our progress in this work is a fundamental understanding of gene expression in the parasite and this has limited our ability to manipulate the organism. Another obstacle had been the limited number of genes that had been examined in the parasite. Progress in the Plasmodium falciparum genome project and development of new technology has provided an opportunity to overcome these obstacles in the parasite. We have now initiated a project to analyze gene expression in Plasmodium falciparum using the newly developed method of Serial Analysis of Gene Expression. This work is being done in close collaboration with Dr. Keith Martin, WRAIR.

Background

The Plasmodium falciparum Genome Project has opened new approaches to drug target identification and through a functional analysis of whole genome expression, new drug targets will be identified. The overall goals of this work are to use the knowledge derived from understanding the profile and mechanism of gene expression to identify novel targets for drug development. Approximately 60% of the predicted genes in the

Plasmodium falciparum genome do not yet have an identified function and among these will be genes critical for parasite survival and function. By using a functional genomics approach to analysis, we hope to identify new classes of genes which we cannot necessarily predict based on homologies with genes identified in other systems or predicted based on common metabolic pathways. These pathways may also help us in identifying new targets for drug development

Genomes to Drugs – Opportunities to discover new drug targets

In order to achieve this, a more fundamental understanding of parasite biology is needed. Great progress has been made in the sequencing of the *Plasmodium falciparum* genome with two chromosomes assembled and annotated (Gardner et al. 1998; Bowman et al. 1999) and high-throughput sequencing analysis complete for over 80% of the genome at the three genome sequencing centers, The Institute for Genome Research (TIGR), the Sanger Centre and the Stanford Genome Center. Relating genomic sequence to function and ultimately malarial biology is the next logical step. One approach involves investigating transcriptional profiles in the parasite at the level of the entire genome. By understanding the network of genes expressed by an organism, complex and interrelated cell functions can begin to be unraveled. Much previous work has focused on single genes, yet many biological processes are the result of interactions of multiple genes and gene products. Such global transcriptional studies are a first step to identifying participants in such complex processes as response to drug treatment. By investigating gene expression on a genome-wide scale, we hope to discover key features of the parasite's biology and discover new targets that could lead to novel drug development.

Serial analysis of gene expression (SAGE) is particularly well suited for malarial systems, as little is known about gene expression and many of the genes identified in the sequencing project are of unknown function. SAGE is an extremely powerful tool with which to simultaneously and quantitatively analyze mRNA transcript profiles from a given cell population, allowing for the discovery of new genes. For the first time, it is now possible to examine the response of all of the genes to stimuli such as drug treatment. Techniques have been developed in other systems and adapted to malaria including differential display (Liang et al. 1992) (Thelu et al. 1994), microarray analysis (Lashkari et al. 1997) and Serial Analysis of Gene Expression, SAGE (Velculescu et al. 1995) (Hayward et al. 2000).

Serial Analysis of Genes Expression

SAGE is well suited to an organism whose genome is not completely annotated and provides an open platform for new gene discovery. SAGE allows the discovery of new genes, as well as the detection of low abundant transcripts by qualitatively and quantitatively analyzing thousands of transcripts in a given cell population at the same time. The technique is based on three experimentally confirmed principles (Velculescu et al. 1995): a) a short (10bp) tag from a defined position within a transcript can uniquely identify a gene. This is reasoned by the fact that the maximum number of possible tag sequences, assuming a random nucleotide distribution (4^{10}= 1,048,576), is far greater than

the number of estimated genes in most organisms; b) concatenation of several tags into a single molecule allows for efficient sequencing and acquisition of data; and c) expression patterns of induced genes are accurately represented by the abundance of their corresponding tags. As such, SAGE can achieve levels of transcript profiling that have not been approached by differential display, subtractive hybridization and EST (expressed sequence tag) technologies (Carulli et al. 1998).

SAGE has been successfully applied in a number of different systems; for example, it has been used to a) characterize the entire repertoire of expressed transcripts in yeast (Velculescu et al. 1997); b) identify p53 regulated genes (Madden et al. 1997) (Polyak et al. 1997); c) compare differential gene expression between normal human and cancer cells (Zhang et al. 1997; Hibi et al. 1998; Hibi et al. 1999; Lal et al. 1999); and d) profile gene expression in rice seedlings (Matsumura et al. 1999). In summary, SAGE lends itself as an extremely efficient tool for qualitative monitoring of global gene expression.

Global gene expression responses to drug treatment

The recent availability of complete genome sequences and methodologies to scan whole genomes has allowed investigators to ask questions about the global response of cells to various stimuli (Schena et al. 1995; Schena et al. 1996; Heller et al. 1997; Velculescu et al. 1997; Schena et al. 1998). Much of the initial work was done in *Saccharomyces cerevisiae* and has led to the surprising observation that over 100 genes change in expression levels when cells encounter toxic drugs or nutrient levels change (DeRisi et al. 1997; DeRisi et al. 2000). These results imply that response to drug treatment involves the interaction of several gene products and several pathways may exist by which the cell can resist the toxic effects of the drug. Over time, a particular pathway may predominate in resistant cells, but the expression of other proteins may continue to play an important role. In yeast, more than 20 genes are turned on after treatment with phorbol ester implying specific transcriptional activation and thus opening a new avenue for the development of interventions. A specific aim of this project is to investigate the global response of *Plasmodium falciparum* to treatment with antimalarial drugs. This research will be accomplished using the methods of Sequential Analysis of Gene Expression (SAGE) (Velculescu et al. 1995)

A second and related question concerns the mechanism by which parasites are killed. Does treatment with toxic drugs result in a unique gene expression pattern or do cells during the course of the response to drug treatment express a common set of genes? Little is known about the events that lead to cell death in parasites. Is the response to each drug or immune mediator unique or is there a common cell-death pathway? In higher eukaryotes, specific pathways are involved in cell death, termed apoptosis, which can be stimulated by many different events including the treatment of cells with toxic drugs. The requirement of programmed and orderly cell death during the development of multicellular organisms is thought to be the origin of the apoptotic pathway. In unicellular organisms such as *Plasmodium falciparum*, whether such a pathway exists remains an open question. Only a single publication in the literature provides evidence for DNA fragmentation after chloroquine treatment, consistent with an apoptotic pathway (Picot et al. 1997). The goal of this work is to explore the parasite's response to several

different toxic compounds, including immune mediators, and examine the gene expression profile using the approach of whole genome scanning.

One of the overall goals of this work is to use the knowledge derived from understanding the mechanisms and networks that control gene expression to identify novel targets for drug development. Approximately 60% of the predict genes in the *Plasmodium falciparum* genome do not yet have an identified function and among these will be genes critical for parasite survival and function. By using a functional genomics approach to analysis, we hope to identify new classes of genes which we cannot necessarily predict based on homologies with genes identified in other systems or predicted based on common metabolic pathways. At least one class of genes will be involved with the regulation of gene expression and these may prove to be unique to *Plasmodium falciparum* and provide new insights into novel aspects parasite biology. These pathways may also help us in identifying new targets for drug and vaccine development.

Experimental Design and Methods

We have been successful in establishing the SAGE technology for *Plasmodium falciparum*. Our results are summarized below and provided in greater detail in the Appendix in Munasinghe et al, submitted and Patankar et al, submitted). We have presented results demonstrating the feasibility of the SAGE methodology applied to the *Plasmodium falciparum* asexual stage parasite system. This sets the stage for examining global gene expression profiles under a variety of conditions and will lead both to the identification of networks of genes coordinately regulated and to the identification of new genes and pathways critical for parasite survival. We have also presented results demonstrating our ability to functionally analyze cis-elements hypothesized to be important for gene regulation. This approach will be critical for the analysis of the gene expression networks identified by the SAGE analysis. In the proposed work, we will extend our SAGE analysis to analyze the parasites under conditions of biological relevance and then use that information to identify genes or groups of genes for further functional analysis.

Development and optimization of the Serial Analysis of Gene Expression (SAGE) technology for *Plasmodium falciparum*

We have demonstrated in the preliminary data and accompanying detailed manuscript the feasibility of applying the SAGE technology to *Plasmodium falciparum*. Under this proposal we plan to use this technology to analyze differential global gene expression under different growth conditions. Additional development of the technology including the development of bioinformatics support for the technology and optimizing the technical aspects of the methodology. The results of our work are summarized below. Additional detailed information can be found in the preprint and manuscript (Munasinghe et al, 2001, Patankar et al, submitted).

Genes expressed by *Plasmodium falciparum* asexual stage parasites

The first set of experiments is to analyze the genes that are expressed by the asexual stage parasites using the SAGE methodology. *Plasmodium falciparum* (3D7) parasites were synchronized and harvested at the trophozoite stage (70% trophozoites) and SAGE analysis. Trophozoites were chosen as the first target for SAGE analysis because of the large body of evidence that in the asexual blood stages, the majority of RNA synthesis occurs during the trophozoite stage. In subsequent experiments, synchronized parasites will be isolated at different stages of asexual blood cycle and expression profiles will be compared. In the initial experiments, we analyzed approximately 7000 individual tags and determined their abundance in the tag population (see Table 1, NB: we only reported the abundance of those tags present at 2 or greater copies.).

As can be seen from this data, only a small percentage of mRNAs are present in high abundance; only 11 genes are found in the highest abundance classes, while greater than 80% of the tags are present at 10 to 50 times lower levels. In addition, those tags present at 2 copies or more in the SAGE library correspond to 1047 genes, or approximately 15% of the total predicted genes in the parasite genome. If the single tag data is included, then trophozoites express approximately 3500 genes, or about 50% of the total predicted genes. This data is quite consistent with other systems such as *Saccharomyces cerevisiae* where about 3500 of the predicted 6500 genes are expressed during vegetative growth with less than 1000 genes expressed at high abundance. This implies that a significant number of genes are not expressed during trophozoite stage of the parasite, although, there may be some transcripts that do not contain an NlaIII site or which are expressed at very low levels or undergo very high turnover.

Table 1.

Percentage Frequency*	**Total number of tags**	**Total number of genes**	**Matches to NCBI *P.falciparum* database**
>2.2	87 (2.4%)	1 (0.1%)	1 (0.1%)
1.1-2.2	90 (2.5%)	2 (0.2%)	2 (0.2%)
0.55-1.0	226 (6.3%)	8 (0.8%)	7 (0.7%)
0.28-0.53	370 (10%)	30 (2.8%)	25 (2.4%)
0.14-0.25	632 (18%)	105 (10%)	78 (7.4%)
0.06-0.11	2196 (61%)	901 (86%)	ND
Total	**3601** (100%)	**1047** (100%)	**ND**

* The tags are divided into abundance classes according to frequency of appearance among 3601 tags comprising the expression profile of the 3D7 control population. The number of tags matching to an entry in the *P. falciparum* database is listed per abundancy class, and the percentage of hits among 1047 unique tags is given in brackets.

The identity of the genes in the highest abundance classes was assigned using the BLAST analysis procedures of both the Genbank database entries at NCBI and the high

throughput shotgun DNA sequence available at the TIGR, Sanger and Stanford Genome Centers. In addition for a subset of the analyses, we used those sequences assembled by Jessica Kissinger and David Roos at the University of Pennsylvania. As can be seen in Table 2 below, several that are expressed genes include

Highly expressed genes		
Tag	**% Abundance**	**Gene description**
TCAGGCGTTA	1.3	cytochrome oxidase (mitochondrial-gene)
GAAGTCGAAA	0.45	5.8S ribosomal RNA
ATTTGAAGCA	0.42	Rhop H3
GTAGTTGACA	0.36	hypothetical protein
CTAAAGCACC	0.33	ras-related nuclear protein
TTGAAGCTGA	0.28	heat shock protein
CGAGGAAAAA	0.27	serine repeat antigen
AACGACAAGA	0.25	Pfg27/25
CCAAATGATG	0.25	polyubiquitin
TACAGCTGCT	0.21	merozoite surface protein
GGGAAAGCGA	0.19	hypothetical protein
TTGAGGATTC	0.19	rifin
GGAAATAAAG	0.18	unknown protein

Table 2: Highly expressed genes in the 3D7 control SAGE library. Tag represents the 10bp SAGE tag adjacent to the NlaIII site. Gene description details the gene corresponding to a particular tag. Abundance is listed as a percentage of all 6702 tags in the 3D7 control SAGE library.

known as well as several matches to hypothetical or unknown proteins. This unknown and hypothetical proteins are likely to represent highly expressed genes in pathways not yet uncovered by traditional approaches and may represent novel parasite pathways, critical for parasite survival and growth. Such genes may point to new pathways to target for drug development.

Northern Analysis to confirm SAGE data

These initial experiments have demonstrated the feasibility of this approach; however, additional data will be needed to fully validate this method and to use the method for further analysis. First, the abundance level indicated by the SAGE analysis will need to confirmed using other methods. Our first approach was to examine several of the genes using the method of quantitative Northern blot analysis and this analysis confirmed the SAGE data (see Patankar et al, submitted). There are two other potential methods that could be used to compare abundance of mRNA. This includes the use of microarrays and we will collaborate with those groups to compare our SAGE data with their data. In addition, for certain genes where the exact amount of mRNA present in the sample is critical to further experiments, quantitative PCR methods can be developed.

Definition of the *Plasmodium falciparum* Transcriptome

A second major aspect of this specific aim is to create a new SAGE tag library in a second experiment using *P. falciparum* 3D7 trophozoites. This will allow us to

determine the reproducibility of the tag library and will also give us additional tags to include in the analysis. An empiric method was developed by Velculescu and coworkers (Velculescu et al. 1995) to determine the minimum number of tags necessary to describe what they termed the "Yeast Transcriptome". They measured the number of new tags obtained as a function of the number of tags analyzed, and determined that a library of 15,000 tags had reached saturation in terms of the acquisition of new tags. In a similar analysis using our initial data, we have determined that for the high abundance class of tags, a library of 2000 tags was adequate for these tags to be identified and to be sorted into the high abundance class. If we analyze total tags as above, then we have not yet reached saturation or plateau level in the acquisition of new tags and predict based on our data and data from the yeast system, that we will need approximately 15,000 tags from a single library. One of our first goals will be to generate such a library to serve as the basis of comparison for all of the other work. Results from this library will be posted on the web, either through our own website and/or through the *Plasmodium falciparum* Genome Project database.

Once the SAGE library has been completed for the trophozoite stage, we will make libraries for the ring and schizont stages of the parasite life cycle. Each of these libraries will be made using synchronized parasites as described in the preliminary data and quantitative analysis of Giemsa stained thin blood smears will be used to assess the purity of the preparations.

Annotation of SAGE tag data

One of the major goals of this work is to understand the networks or groups of genes that are coordinately expressed by the parasite. In the initial experiments, we will determine all of the genes expressed during the trophozoite stage. This should give us insights into metabolic pathways, expression of surface molecules and identify several unknown genes to focus on for further analysis. In our preliminary data we have begun this analysis for the tags expressed in the highest abundance classes (shown in Table 2). We will continue this analysis with the remainder of the tags and more importantly with the larger 3D7 SAGE library. This process will require some additional software development since the majority of the *Plasmodium falciparum* genome sequence is not fully annotated. The work done under this specific aim will contribute significantly to annotation of library with regard to identifying genes with regard to their expression profile. One advantage of this SAGE library is that the once the data is collected, it will continue to be useful in identifying expressed genes as more of the genome is annotated.

High Abundance Class mRNAs

Another important outcome of this work is the identification of key metabolic pathways in the parasite. For example, the most abundant tag corresponds to a mitochondrial gene, cytochrome oxidase. Consistent with this, Vaidya's group (Srivastava et al. 1999; Srivastava et al. 1999) has identified the target of atovaquone, the newest antimalarial drug to be developed, as mitochondrial cytochromes. The baseline data provided by the 3D7 trophozoite stage SAGE library will provide us with a description of all of the genes expressed by the parasite in this metabolically active stage

of the parasite life cycle. In addition, it will give us greater insight into which metabolic pathways are likely most active in the various stages of the parasite life cycle.

Global Gene Expression Following Drug Treatment

One of the most powerful approaches to understanding key pathways in parasite biology is to examine the changes in gene expression under different conditions. Of primary interest to our laboratory is the effect of drug treatment on parasites. This both give us insights into putative mechanisms of action and the potential of identifying new targets in existing pathways. To initiate this work, we have treated *Plasmodium falciparum* parasites with chloroquine under conditions that will kill the parasite over the course of a 48-hour treatment.

5.2 Analysis of Gene Expression in Plasmodium falciparum

As the work on the global analysis of gene expression is ongoing, we have continued our efforts to understand gene expression at the level of the individual gene in order to develop methods to modify expression using molecular genetics. Again, with the additional information provided by the genome project, we have been able to make excellent progress on this work and have made a preliminary observation which indicate that we may be able to readily identify cis-acting elements which control gene expression through an analysis of comparative genomics.

5' Untranslated Region sequence variation in strains of *Plasmodium falciparum*.

Transcriptional regulation has not been well defined in *Plasmodium falciparum*. A review of studies that have looked at the regulation of different genes indicate that regulation in the parasite may be different from the classic model of regulation in eukaryotes. While the mechanism may be different, it is likely that transcriptional regulation plays an important role in genes such as *pfmdr1*. This is supported by the evidence in yeast, where *PDR* genes are transcriptionally regulated. Thus in an effort to further characterize the role of *pfmdr1* in drug resistance, we have begun to map the 5'untranslated region of the gene.

A contig on chromosome 5 containing 3D7 strain *pfmdr1* coding and noncoding sequence has been extracted from the *Plasmodium falciparum* genome database. The coding region of this contig has 100% identity with the coding sequence of the D10 strain *pfmdr1* clone in Genbank (gi:9935). However, the 5'UTR of the genes only has 60% identity. This reduction in identity occurs in a pattern of complete homology, followed by minor variation, followed by major variation at the most upstream end of the 3D7 and D10 sequence. This pattern may be indicative of selective pressure in the gene.

We were interested to see if this pattern of increasing variation was present in other *P.falciparum* genes. A comparison of upstream sequence of the 3D7 and T9/96 (gi: 160127) *P. falciparum* calmodulin genes revealed a similar pattern. Of course, the possibility of cloning artifacts and sequencing errors must be ruled out before the

significance of this result can be determined. Both sets of sequences were aligned utilizing the Clustal W alignment tool. They were anchored at the putative translational start site, and matched in length to minimize gaps. 3D7, the reference strain, is highlighted in bold for both alignments.

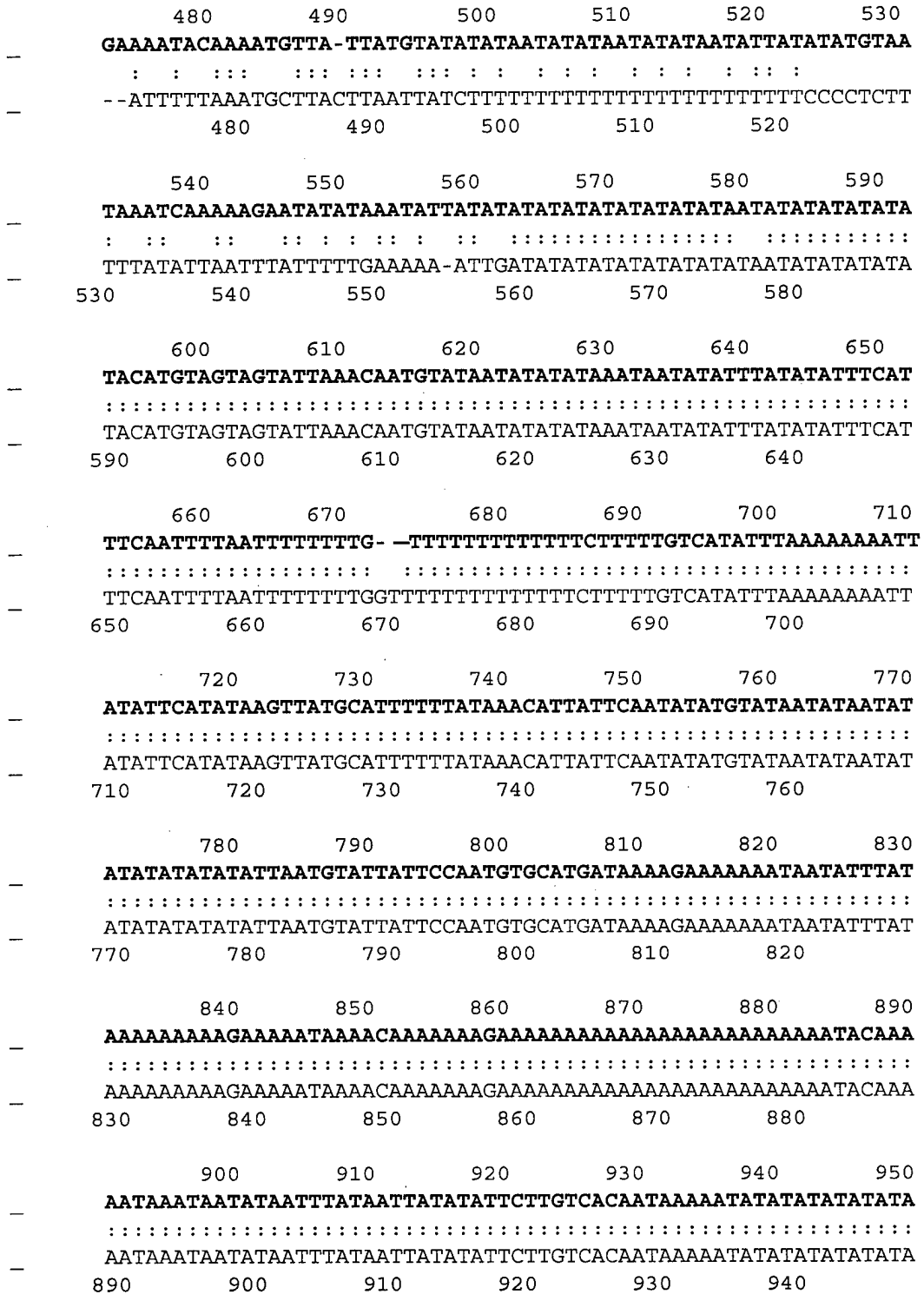
SEQUENCE ALIGNMENT OF 3D7 AND D10 *pfmdr1* 5'UTR.

```
80.5% identity in 503 nt overlap; score: 1217

      10        20        30        40        50        60
      TAAAATATACATAATTAAATATAAAAATGACATTATATTTTTGTTAAATTATACAGAAGA
      ::::   :  ::::  :::::::: ::: :     : :::::  ::::: ::   :   :
      TAAATCTTTTATAA--AAATATAATAATTAATAATTTTTTTTAATAAATAATTTTGTTTA
           10        20        30        40        50        60

      70        80        90       100       110       120
      AAAAAAAAAAAAAAAAAAAATAGAAGTAAATTGTATAGAATTATTTNTTTATTAATATTAT
      ::  :: :: :   ::   ::  : : : :: :  : : :: :::.: ::::  :::
      AATTAATAATATGTAATTTTATTATTTATTTATTAACATTTTTTTTTATATTT-TATATA
             70        80        90       100       110       120

     130       140       150       160       170       180
      TATTTTATTTTGAATAA--AACTATTTTNGTATCTAATAATAAATA-TAATAACACATAT
       : :  :: :  :::::  :: : : : .: :: ::: :::: :: :::: :  :::
      AAATACATATATAATAACTAAATTTATGCGCATATAAAAATATCTAATAATTTTAATTAT
             130       140       150       160       170       180

       190       200       210       220       230       240
      ATATATATATATAT-ATATATATATATATTATT---TNANTTATTATTATATTTTTTTTT
      :::::: ::::::: ::: ::::: ::::::    :.:. ::::::::::::: : :: :
      ATATAT-TATATATTATACATATAATTATTATAACGTTATATATTATTATATTATATTAT
             190       200       210       220       230

           250       260       270       280       290       300
      TTATTATTTTTTTTGTCATTGTGTAAATATATAAATATATATANTATATATATTATTA
      : :::::::::::::::::::::::::::::::::::::::::.:::::::::::::::
      T-ATTATTTTTTTTGTCATTGTGTAAATATATAAATATATATATTATATATATTATTA
     240       250       260       270       280       290

           310       320       330       340           350
      TTTCAACATTGTTTATATATATATATATATATATATATAT----TTATATTTATATATTG
      :::::::::::::::::::::::::::::::::::::::    ::::::::::::::::
      TTTCAACATTGTTTATATATATATATATATATATATATATATTTATATTTATATATTG
     300       310       320       330       340       350

     360       370       380       390       400       410
      ATATATGTGTACATAGCTTATTTCATTTATAAGATTTAGATTTTGTTTTTAATATTATAT
      ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
      ATATATGTGTACATAGCTTATTTCATTTATAAGATTTAGATTTTGTTTTTAATATTATAT
     360       370       380       390       400       410

     420       430       440       450       460       470
      AATTTTGTTTGTTACAAATTAATTAATTATTTCTTTATTTCTTTATTTTATTTACATTTT
      ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
      AATTTTGTTTGTTACAAATTAATTAATTATTTCTTTATTTCTTTATTTTATTTACATTTT
     420       430       440       450       460       470
```

Sequence alignment of 3D7 (TOP) and T9/96 (BOTTOM) Calmodulin 5'UTR.

76.2% identity in 1015 nt overlap; score: 2163

```
      60        70        80        90       100       110
      AAATATTATTTATAACAAGAGAAAAGGCAGAAACAAAATAA-ATTATAATAAAAAACACA
      ::  ::: : :: :: ::   ::::   :   : : :::: ::: :: :  :  : : :
      AACCATTTTGTAAAAAAAATTAAAATATATTTATATAATATTATTTTATTTTATTATATA
            80        90       100       110       120       130

     120       130       140       150       160       170
      TTTTTTTATATTTGTATGAATATATTTTTTGTTATGCCTAAAAAAAAAATAGGATTATC-A
      :: : :::: ::: : :  :: : : :::: :: :  ::: :::          :::::
      TTATATTATTTTTATTTTTATTTTTATTTTTTTTTCTCTACAAATT-------TTATCTA
           140       150       160       170               180

      180       190       200       210       220           230
      TATTTTTATATAAAATGTAAGGATTTCAAAATATATATAATTT---TTTAAAATAACAAA
      :   :::::   :::  ::   ::::: ::  ::: ::::::    : : :: : : : :
      TTGGTTTATTATAAAAATATCTATTTCTAATAATAAATAATTAAGATATCAATTTATAGA
        190       200       210       220       230       240

        240       250       260       270       280       290
      AAGGGAACATTTTTTTTTTTTTTTTTAACATTTTCATGCCACGTTGACAAGAATTTTTAA
      ::   :: :: :  :: : :  :::::   ::::      :   : :::   ::  :::
      AACAAAATATATACTTGTATAATTTTATTTTTTTATATAAATCATTACATATATAATTAT
        250       260       270       280       290       300

        300       310       320       330       340       350
      AAAATCCATTAAATTAAAAATAACTTTTTTATTTATTTAAATAAGATATTCAAATAAGGA
      : :::   ::   : : : ::::     :::: ::::     :: ::::  ::: :: :
      ACAATATTTTTTCTAAGAGATAA-----TTATATATT-----AATATATATAAAAAAAGG
        310       320             330             340       350

        360       370       380       390       400       410
      TATTTATTAATTAGCTCGCAAATGGCCAAATAAGAAATATAATATAATATATTATTATAT
      : ::: ::  ::   :    : :       ::    : : ::   :::::: ::::: :
      TGTTTTTTTTTTTTTTTTTTTATTTTT---ATTTTTATTTTATGGTAATATTTTATTTTCC
        360       370       380           390       400       410

        420       430       440       450       460       470
      ATATTATATATATATATATATATAAATATATTTATAATAATA-ATATAAATAAAGTATAT
       :::: :::: :: :::::   :  :::: :   ::::  :: :::: :  ::  :::::
      TTATTTTATAAAT-TATATTAGTTTATATGTGATTAATTTTATATATTATCAATTTATAT
           420       430       440       450       460       470
```


480 490 500 510 520 530
GAAAATACAAAATGTTA-TTATGTATATATAATATATAATATATAATATTATATATGTAA
--ATTTTTAAATGCTTACTTAATTATCTTTTTTTTTTTTTTTTTTTTTTTTCCCCTCTT
480 490 500 510 520
540 550 560 570 580 590
TAAATCAAAAAGAATATATAAATATTATATATATATATATATATATAATATATATATATA
TTTATATTAATTTATTTTTGAAAAA-ATTGATATATATATATATATATAATATATATATA
530 540 550 560 570 580
600 610 620 630 640 650
TACATGTAGTAGTATTAAACAATGTATAATATATATAAATAATATATTTATATATTTCAT
TACATGTAGTAGTATTAAACAATGTATAATATATATAAATAATATATTTATATATTTCAT
590 600 610 620 630 640
660 670 680 690 700 710
TTCAATTTTAATTTTTTTTG- —TTTTTTTTTTTTTCTTTTTGTCATATTTAAAAAAAATT
TTCAATTTTAATTTTTTTTGGTTTTTTTTTTTTTTCTTTTTGTCATATTTAAAAAAAATT
650 660 670 680 690 700
720 730 740 750 760 770
ATATTCATATAAGTTATGCATTTTTTATAAACATTATTCAATATATGTATAATATAATAT
ATATTCATATAAGTTATGCATTTTTTATAAACATTATTCAATATATGTATAATATAATAT
710 720 730 740 750 760
780 790 800 810 820 830
ATATATATATATTAATGTATTATTCCAATGTGCATGATAAAAGAAAAAAATAATATTTAT
ATATATATATATTAATGTATTATTCCAATGTGCATGATAAAAGAAAAAAATAATATTTAT
770 780 790 800 810 820
840 850 860 870 880 890
AAAAAAAAAGAAAAATAAAACAAAAAAAGAAAAAAAAAAAAAAAAAAAAAAAAATACAAA
AAAAAAAAAGAAAAATAAAACAAAAAAAGAAAAAAAAAAAAAAAAAAAAAAAAATACAAA
830 840 850 860 870 880
900 910 920 930 940 950
AATAAATAATATAATTTATAATTATATATTCTTGTCACAATAAAAATATATATATATATA
AATAAATAATATAATTTATAATTATATATTCTTGTCACAATAAAAATATATATATATATA
890 900 910 920 930 940

Functional analysis of cis-elements in Plasmodium: Basal transcriptional element identified

The goal of this work is to identify and characterize cis-elements that regulate gene expression in the malaria parasite. We have used the model system developed under this funding, namely transfection of Plasmodium gallinaceum zygotes, for these experiments. It has the advantage of being a robust system which readily allows functional testing of cis-elements including analysis of mutated sequences and subsequent biochemical characterization. The first manuscript for this work has been accepted for publication and a second is in the final stages of preparation. The work is summarized below. The technology described here will be applied to analysis of P. falciparum putative promoter elements (see above).

The malaria parasite undergoes a complex developmental process through its life cycle. This includes an asexual intraerythrocytic cycle in the vertebrate host, and a sexual cycle that commences with gametogenesis in the vertebrate host and subsequent fertilization and maturation in the mosquito vector. Regulation at the transcriptional and post-transcriptional levels is no doubt important for the temporal expression of genes required at each stage of development. Present understanding of the *cis*-elements important for transcriptional control in *Plasmodium* is severely restricted. Sequence analysis of 5' flanking regions of *Plasmodium* genes reveal the presence of sequences with homology to known eukaryotic control elements, for example see [1, 2]; however, the functional significance of these sequences in *Plasmodium* has not been demonstrated. The intergenic region in *Plasmodium spp.* Is particularly AT-rich, even within the context of the AT-biased (~80%) genome [3], such that even the identification of TATA-like elements, and assays to determine their utility and importance, becomes difficult. A growing but limited number of functional analyses of promoter regions of *Plasmodium* genes have been published, many of which shed light on regions that are necessary for efficient expression [2, 4]. However, only a few studies to date have identified specific sequences, short of transcriptional start sites, that appear to be important for gene expression [4-6]. Due to the small numbers, and the fact that these genes are expressed at different stages in the parasite life cycle, no consensus or common sequences could be established. Clearly, much more can be learned about aspects of basal transcription as well as stage-specific control of gene expression in the malaria parasite.

Pgs28 is expressed abundantly on the surface of mosquito stages of the avian parasite, *Plasmodium gallinaceum*. Pgs28 belongs to the family of Pxs proteins, which includes the *P. berghei* homolog Pbs21 and the *P. falciparum* homolog Pfs25. These proteins contain a series of EGF-like domains that may serve a function in cell □ignaling or in adhesion [7, 8]. Pgs28, Pfs25 and Pbs21 had been identified as targets for transmission blocking antibodies [9-12]. Transcripts of *pbs21* had been observed in female gametocytes and gametes, as well as zygotes and ookinetes, and the *pfs25* promoter appears to be specifically active in mosquito stage parasites, supporting the notion that the genes encoding this protein family are activated specifically during the sexual stages [5, 13, 14]. Since Pbs21 is initially expressed on the surface of zygote stage

parasites, additional post-transcriptional control is exerted by the parasite to regulate Pbs21 expression. We are interested in investigating *pgs28* gene expression to further understand transcriptional regulation in *Plasmodium spp*. And as a step towards understanding the control of sexual development in *P. gallinaceum*. In this report, we present a functional analysis of the 5' flanking region of *pgs28*, using firefly luciferase as a reporter, by which we identified two regions that are required for *pgs28 trans*-gene expression. Furthermore, using Northern analysis, we define the 5' limit of the *pgs28* transcript and demonstrate that *pgs28* transcripts are present during the zygote stage.

The 5' and 3' flanking sequence of *pgs28*, together with an in-frame insertion of the luciferase reporter, were previously cloned into pBS (*pgs28.1LUC)* [15]. In this study, the *pgs28-luc* chimera, containing *pgs28* 5' flanking sequence, the *pgs28-luc* fusion gene, and about 720 bp of 3' flanking sequence, from *pgs28.1LUC* was cloned into the *Hind*III site of pBS to create *BSpgs28-LUC.* The 1871 bp 5' flanking sequence of *pgs28* has been determined and deposited in GenBank. (The sequence and characterization of the 3' region was recently published [16].) Expression from *BSpgs28-LUC* was confirmed by immunofluorescent antibody staining and immuno-electron microscopy [17] and also by luciferase assays performed 24 or 48 hrs post-transfection (see below). High expression levels up to the order of 10^6 light units were obtained, offering a sensitive system for determining changes in expression levels.

In order to determine the sequence requirements for *pgs28* expression, a series of 5' deletion mutants was created either by exonuclease digestion of linearized *BSpgs28-LUC* plasmid, or by PCR mutagenesis. Deletions of 790bp (FP1081), 1131bp (FP740), 1358bp (FP513), 1407 (FP464), 1485bp (FP386), 1538 bp (FP333), 1584bp (FP287), 1631 bp (FP240) and 1905bp (FP+34) from *BSpgs28-LUC* were obtained (Fig. 1A). Additionally, an internal deletion mutant □376-316, which lacks the specified sequences, was created by inverse PCR. To assess the contribution of the deleted sequences to *pgs28-luc* transgene expression, these plasmids were transfected into sexual stage parasites as previously described [15]. Luciferase activity was assayed after 24 or 48 hours. To control for transfection efficiency, a second plasmid, *pgs28-GUS*, was co-transfected and luciferase light units normalized to GUS fluorescence units.

Transfection using FP1081 demonstrated that expression of the *pgs28-luciferase* fusion gene did not decrease significantly when the 5' most 790 bp were deleted from the parent plasmid (Fig. 1B). However, luciferase expression from FP740, where an additional 340 bp has been removed, was reduced by more than 40%. A modest decrease in promoter efficiency was observed with the removal of the next 227 bp (FP513). Further deletions of up to 180 bp (FP464, FP386, FP333) did not seem to significantly affect expression when compared to FP513. Interestingly, FP287, containing a deletion of 46 bp 3' of FP333, had less than 5% activity compared to the full-length construct. Furthermore, the internal deletion mutant □376-316 was also severely affected, having only 6.6% activity. As expected, a deletion that encompasses part of the *pgs28* open reading frame (FP+34) abolished luciferase expression. The mutant FP240 had equivalent activity to this construct, suggesting that important elements necessary for transcription

and possibly translation have been removed. Taken together, results of the 5' deletion analysis suggest that the minimal sequence necessary for *pgs28* transgene expression consists of the 333 bp upstream of the translational start site. Moreover, a 17 base-pair sequence, TACCATTTGTACAGACAG, between –333 and –316, appears to be crucial, since *pgs28* expression was essentially abrogated in a 5' deletion mutant and an internal deletion mutant that lack these sequences. We suggest that the proximal site corresponds to the basal promoter or initiator element, as indicated in the following section. We also suggest that positive regulatory elements lie between –1081 and –740, and perhaps within –740 and –513. This distal region likely contains an enhancer element(s) that contributes to *pgs28* promoter efficiency. Thus transcriptional elements that control *pgs28* appear to be bipartite, as in eukaryotic promoters and other *Plasmodium* genes that have been analyzed.

We used Northern analysis as a preliminary step to map the transcriptional start site of *pgs28*, and to determine whether the temporal pattern of *pgs28* transcription paralleled that of its murine homolog *pbs21*. RNA was isolated from newly formed zygotes collected after exflagellation, and from gametes. As seen in Fig.2B, an intense signal appeared at a position corresponding to a message of about 1.4 kb in both zygote and gamete when probed with BBm600 (lanes 1 and 2), which extends from –381 in the 5' flanking region to within the *pgs28* coding sequence. A *pgs28* message of 1.5 kb has previously been reported by Duffy and colleagues [11]. Thus, while Pgs28 expression is most abundant on ookinete surfaces, *pgs28* transcript can be seen as early as the zygote stage. This is in agreement with transfection studies in our laboratory using the *BSpgs28* construct, as well as a *pgs28-GFP* fusion, that demonstrated Pgs28 expression on the surface of zygotes [17].

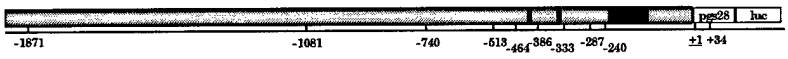
Recently, the polyadenylation signal of *pgs28* was mapped to approximately 425bp downstream of the stop codon, with an estimated poly(dA) tail of at least 20 nucleotides [16]. Given that the coding sequence of *pgs28* is 666bp, the transcription initiation site of *pgs28* would lie approximately between –390 and –290bp. In agreement with this estimation, only the probe BV142, encompassing the sequence from –381 to –240, hybridized to the *pgs28* transcript (Fig. 2B, lane 7), while probes corresponding to sequences further upstream failed to hybridize to *pgs28* mRNA (lanes 3-6). Thus, these studies establish the 5' limit of the *pgs28* transcript at –381 bp upstream from the translational start site. 5' deletion analysis suggests that the transcriptional start site is likely to be downstream of –333bp. Experiments to determine the precise 5' end of the *pgs28* transcript will resolve this aspect of *pgs28* transcription.

The 5' flanking sequence of *pgs28* had been inspected for homology to other eukaryotic transcriptional regulatory elements. The highly AT-rich region between –1081 and –520, typical of intergenic regions of *Plasmodium spp.*, does not contain sequences that are analogous to known eukaryotic regulatory elements. Two GTAAT sequences, demonstrated to be important for *GBP130* expression [6], can be found in this region. Whether an element associated with an enhancer of an asexual stage gene in *P.*

falciparum is important for expression of *pgs28*, a sexual stage specific gene, can only be determined by experimental means.

Sequences downstream of –520 have also been examined. Within this region are two putative TATA elements TAAAAAGAATAA and TATAAATGTTT, centered at –434bp and –360bp respectively from the start codon. Since these sequences can be deleted from the reporter constructs (FP386 and FP333) without drastically affecting expression, they are not likely to be important for *pgs28* expression. This again illustrates that sequence analogy to eukaryotic promoter elements does not necessarily imply functional significance in *Plasmodium* genes. Inspection of the presumed 5' UTR reveals a T-rich stretch, constituting up to 74% of the bases between 130bp and 242bp. A series of five 8-base pair inverse repeat elements (*TTTA**TTTT**ATTT*) could be identified within this sequence. Further examination of this region uncovers 3 direct repeats of 27bp to 29bp in length. Whether these sequences have functions at a post-transcriptional step to enhance *pgs28* expression awaits further experimentation. Recently, transfection studies of *pfs25* promoter constructs into *P. gallinaceum* ookinetes, and mobility shift assays using *P. gallinaceum* ookinete nuclear extracts, suggest that the sequence AAGGAATA, found at –403 to –396 and –483 to –476 from the initiation codon in *pfs25*, interacts with a nuclear factor and is important for expression of *pfs25* [5]. A similar sequence, AAGAATAA, is found at –354 and –347 in *pgs28*, within the putative proximal TATA sequence. Again, the transfection studies reported here suggest that this sequence in *pgs28* can be deleted without severely affecting *pgs28* transgene expression. This suggests that the nuclear factor PAF-1 [5] is not involved in *pgs28* transcription, and/or that it has a stringent sequence requirement that the AAGAATTT sequence in *pgs28* does not satisfy. Even though *pgs28* and *pfs25* belong to the same family, and possess similar expression profiles during the parasite life cycle, they may not necessarily be controlled by the same evolutionarily conserved factors. Nonetheless, given the close evolutionary relationship between *P. gallinaceum* and *P. falciparum*, it would be of great interest to determine whether the 17 bp upstream sequence in *pgs28* between –333 and –316 would be able to functionally replace the *pfs25* sequence, and vice versa.

A.


pgs28
luc
-1871
-1081
-740
-513
-464
-386
-333
-287
-240
±1
+34

B.

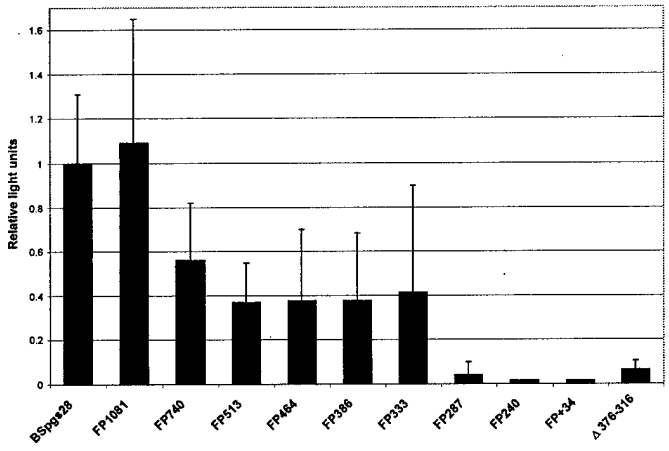

Relative light units
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
BSpgs28
FP1081
FP740
FP513
FP464
FP386
FP333
FP287
FP240
FP+34
Δ376:316

Fig. 1A. Schematic of the *pgs28* 5' flanking sequence.
The 5' flanking sequence of *pgs28* cloned into *BSpgs28-LUC* is shown, together with part of the *pgs28* and *luc* coding sequence. To obtain *BSpgs28-LUC, pgs28.1LUC* [15] was digested with *Hind*III and cloned into similarly digested pBluescript KS+ (Strategene).
The bars at approximately -440 and -360 represent two putative TATA boxes. The hatched box downstream of -240 represent the T-rich sequence with internal repeats.
Positions of the 5' deletion mutants are indicated. The numbers refer to the distance in nucleotides away from the start of the coding region (+1).

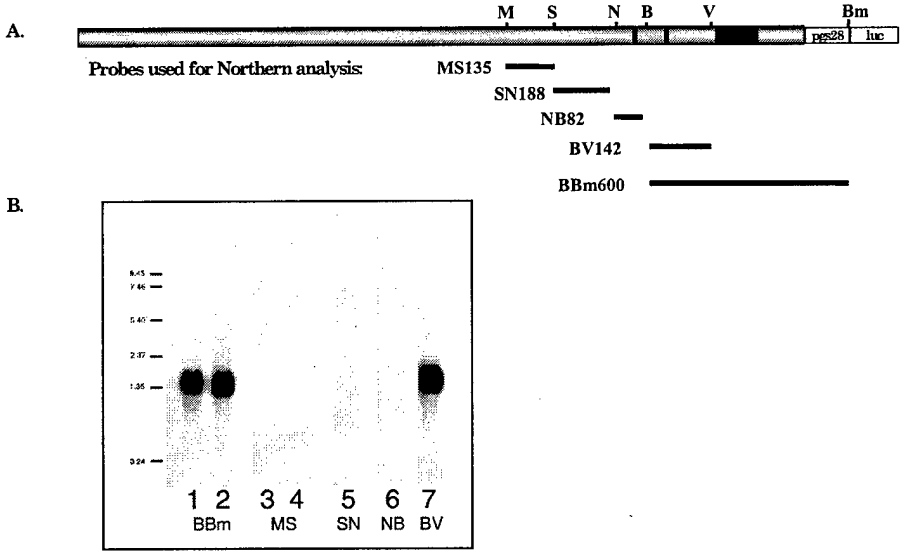
To generate the 5' deletion mutants FP1081, FP464, FP513, FP386 and FP+34, *BSpgs28-LUC* was first digested with *SacI* and *SpeI* (New England Biolabs). The linearized plasmid was digested further with exonuclease III/mung bean nuclease essentially as described by the manufacturer (Strategene). *E. coli* (XL-1 Blue) cells were transformed with ligated products and the sizes of the plasmids obtained were determined by agarose gel electrophoresis. FP464 was generated by recloning the filled-in *Nde*I insert from *BSpgs28-LUC* into *Sma*I digested pBS. FP333, FP287 and FP240 were created by PCR mutagenesis, using FP464 as template, and the upstream primers
5'GAATTCCTGCAGCCCTACCATTTGTACAGAC,
5'GAATTCCTGCAGCCCACTAGCTAAAAGAAATATG, and
5'GAATTCCTGCAGCCCATTTTTATTTAATTTTTC respectively. The *Pst*I site is underlined. The downstream primer, 5'CTAGAGGATAGAATGGCGCCG, containing an internal *Sfo*I site (underlined), was used in all cases and was derived from the *luc* coding region. Purified PCR fragments were digested with *Pst*I and *Sfo*I and cloned into similarly cut FP464 vector backbone.

To generate ? 376-316, primers WFM48 5'CCATTTGTTATTGTATATAAAAAAAAAAAC and WFM20R 5'GATCTTCTTAATCTTTGTAAAAATAACTG, which flank the sequences to be deleted, were used to amplify FP513 that had previously been linearized with *Bgl*II, utilizing the *TaqPlus Long* PCR system (Strategene). 30 cycles of PCR reactions were performed in low salt buffer under the following conditions: 94°C for 1 min, 55°C for 1 min and 72°C for 7 mins. PCR products were treated with *Dpn*I at for 30 mins and further treated with 1? 1 of *Pfu* for 10 cycles and incubation at 37°C for 30 mins. Amplified products were phenol:chloroform extracted and ethanol precipitated, and resuspended. Amplified DNA containing the deletion was allowed to circularize and transformed into *E. coli* (XL-1 Blue) cells. Sequences of all clones were confirmed by DNA sequencing.

B. Luciferase expression from *pgs28* 5' deletion mutants.
Parasites were transfected with the indicated plasmids and *pgs28-GUS*, and luciferase and GUS activity assessed 24 or 48 hrs post-transfection, as described. The construction of *pgs28-GUS*, containing an in-frame insertion of the *? -glucuronidase* gene (Clontech) within *pgs28*, has been described elsewhere. Normalized relative light units and SD are shown. The indicated activity is the average of 3-8 determinations.

Fig. 2A Positions of DNA probes used to map the 5' end of *pgs28* transcripts.
Probes MS135 (–786 to -651), SN188 (-651 to –463), NB82 (-463 to –381), and BV142(-381 to –240) and BBm600 (-381 to +217) were made by digesting an *XbaI* fragment of *BSpgs28-LUC* containing *pgs28* sequences with *MwoI/SwaI, SwaI/NdeI, NdeI/BglII, BglII/VspI and BglII/BamHI* restriction enzyme pairs, respectively. These resulted in fragments of lengths indicated by the numerals in the designations.

B. Determination of size and the 5' end of *pgs28* mRNA from *P. gallinaceum*
RNA was extracted either from zygotes (lanes 1 and 3) or ookinetes (lanes 2, 4-7), fractionated and Northern blotted using standard procedures. Between four and five micrograms total RNA obtained from $3x10^7$ parasites were included per lane. Blots were probed with the indicated DNA fragments, washed and autoradiographed for 24-48 hours.
Lanes 1 and 2, BBm600; lanes 3 and 4, MS135; lane 5, SN188; lane 6, NB82; lane 7, BV142.



Mbacham et al

5.3 **Functional analysis of putative drug resistance and new drug target genes in the heterologous yeast expression system** (for Figures, see Nau et al., 2000 in Appendix)

1. Identification of new drug target genes through complementation analysis in yeast. A single yeast strain with mutations in PDR5/10/SNQ2 has been chosen for this work.

2. Development of new, rapid, high throughput drug screening methods for malaria genes expressed in yeast

This work continues as previously and also includes new initiatives that make use of technology not previously available. The use of the Yeast DNA Microarray in collaboration with the groups of Dr. Maryanne Vahey, Dr. Dennis Kyle and Dr. Keith Martin, WRAIR has greatly facilitated our work and identified new approaches to drug development using this heterologous system. The initial results of this work have been submitted for publication and the ongoing work is summarized below.

(i) Microarray analysis

The majority of the work towards my thesis has involved the analysis of global expression patterns of Saccharomyces cerevisiae when exposed to the antimalarial compound chloroquine. Our laboratory is interested in the mechanisms of drug resistance employed by protozoan parasites with specific interest in the role of membrane transports in resistance. Chloroquine is an important compound for malarial treatment and an understanding of the ways in which organisms respond and develop resistance to this compound is of interest. The choice of yeast as the model system for this analysis is based on several points:

- The global expression response to antimalarial compounds has not been investigated in any organism

- The yeast system provides a unique combination of tools in the form of a complete genomic sequence with approximately 70% annotation of function and microarray technology that allows the simultaneous observation of all of the Open Reading Frames (ORFs) of the yeast genome

- Presence of a network of ATP-binding Cassette (ABC) transporters (pleiotropic drug resistance) with similarity to ABC transporters found in other systems that are involved in Multi-drug resistance phenotypes (see figure 1 for list of yeast ABC transporters)

- The availability of yeast strains produced by functional knock-out that are sensitive to compounds of interest

This type of analysis will not only provide a better understanding of how the Pleiotropic Drug Resistance (PDR) network of yeast, and the corresponding ABC-transporters in the parasite system, functions but will also provide leads to additional mechanisms and homologues in the parasite system.

B. Materials and Experimental Design

This study is utilizing the Affymetrix Gene Chip? Ye6100 yeast chip system. This system is composed of 5 chips, 1 test chip and 4 yeast ORF chips. The test chip is used for quality control and contains representative ORFs from several organisms and a set of spike controls that are also present on the 4 main yeast ORF chips.

The 4 main chips contain probes for approximately 6200 yeast ORF that cover virtually the entire yeast genome. Each probe set consists of ~20 sets of 25mer oligonucleotides that are exact matches to the genomic sequence and corresponding sets of one base mismatches. The mismatch sets provide controls for background and non-specific hybridization. The resolution range for this assay is 0.1-100 mRNA molecules per cell.

The yeast strains used for this study are the PDR functional knockout YHW1052, which has functional disruptions in 3 ABC-transporters, and the parental wild-type strain YPH499. The specific genotypes for these strains are given in table I.

Three treatment points were selected based on the growth curves depicted in figure 6. The three points increase in severity with increase in number: T1 (2hr-1.5mg/ml), T2 (3hr-2.5mg/ml), T3 (4.5hr-2.5mg/ml). The T1 treatment was selected to examine the expression profile at a point on the growth curve just before the two strains diverged from one another with the expectation that expression levels would already have significant differences. The T3 point was selected to examine the profile under extreme drug stress.

C. Results

The gross global expression patterns observed over the three treatment points for each strain are depicted in figures 7 and 8. These graphs show all ORFs that had a differential of 3-fold in expression levels comparing drug treated to control. These ORFs are divided into 12 functional families by their annotations in the Saccharomyces Genome and MIPS Genome Databases. It is interesting to note that although roughly 70% of the yeast database is annotated with either similarity or direct functional data the category of Unknown Function (UNK) still ranks as the top group for expression response to the drug treatment as compared to control. This has also been observed in other microarray studies such as those conducted by the Jelinsky and Samson.

Comparing the two profiles there is a significant difference in expression response between the two strains. The peak in expression response, measured by number of ORFs having a 3-fold differential, for the wild-type parental strain (YPH499) is in T2 while that of the functional knockout (YHW1052) occurs during T3. The majority of the this peak response for the YHW1052 strain is a decrease in expression as compared to the control and the cells appear very unhealthy on visual inspection. It is possible that the expression profile for the functional knockout strain in T3 is largely the result of cellular death processes.

Another indication of the differences in expression profiles between the two strains is the small number of high response ORFs that are in common for the two strains. Both strains have more than 80 ORFs responding with a 6-fold change but only 11 of these are shared. Once again roughly half of these ORFs are of unknown function.

Challenges associated with this type of analysis are data handling and the selection of specific targets for further study. These include the ABC-transporters of the PDR network and two members of the Major Facilitator Super family (MFS) of small molecule transporters.

The PDR transports were selected based on the interests of our lab as a whole and the relation of this network of transporters with several aspects of drug resistance. Observations made by our laboratory and that of Karl Kuckler's, the source of our pdr yeast strains, indicated overlapping function and substrate specificity for members of the pdr network. The current study provides an opportunity to investigate the expression patterns that associated with this phenomenon and to more clearly elucidate the interactions of these transporters.

The MFS transporter SIT1, an iron siderophore transporter, has been selected based on the magnitude of its expression in the two strains and its status as one of the shared ORFs in the expression responses of the strains 2 (Lesuisse et al 1998). YOR273C, the other MFS member, was selected based on the support of our expression data by an independent functional screen performed by Delling et al. in which a yeast genomic S.c. library was screen for conference of resistance to quinoline ring-containing antimalarial compounds (Delling, et al 1998)

In the wild-type strain (YPH499) PDR5 has a small but significant increase compared to control in the treated sample. The PDR5 gene is the member of the PDR network that shows the greatest similarity to the PfMDR1 gene of Plasmodium falciparum. In the case of the functional knockout strain (YHW1052) there are significant increases in three members of the PDR network family, PDR12, PDR15, and YOR1, in response to the removal of PDR5, PDR10, and SNQ2. This observation appears to further support the hypothesis that these transporters have overlapping responses and substrate specificity.

YOR273C shows significant expression in both strains and this expression is

supported by Northern Slot analysis.

D. Summary and Future Directions

- Chloroquine treatment affects the expression of >200 ORFs in each strain
- Gene expression profile is dependent on genetic background; specifically the functional knockouts have a significant impact
- Expression of PDR-related transporters supports the predicted roles and the hypothesis of overlapping response and substrate specificity
- There are several indications for a role of YOR273C in responses to quinoline ring compounds
- Northern Slot analysis confirms the chip data on YOR273C

Northern Slot analysis will be continued to confirm chip data on other ORFs of interest. Further analysis of the array data will be conducted using cluster and temporal approaches in order to discover further associations and patterns. Our data will also be compared with other published and available array data in order to discern general stress responses and other general response phenomenon. Overexpression and Knockout experiments are planned with selected targets. Additional knockout strains for PDR transports are in hand and analysis of these is underway.

In addition we are interested in conducting an additional chip analysis of the original strains with another compound (FK506, fluconazole, ketoconazole, rhodamine6G) as yet to be determined.

Confirmation of Complementation and Mating Phenotype

Recently we demonstrated that expression of *PfMDR1* in yeast deficient for *ste6*, resulted in complementation of the mating phenotype conferred by the native STE6 protein in yeast (Volkman, *et al.*, *PNAS* (95) **92**, 8921]. Ruetz et al. [*PNAS* (96) **93**, 9942] reported complementation of *ste6* with *PfMDR1*, and that expression of *PfMDR1* conferred drug resistance in yeast for quinine, quinacrine, mefloquine and halofantrine. The observation that *PfMDR1* expression conferred drug resistance in yeast is different than our findings that *PfMDR1* expression is associated with increased drug sensitivity in yeast. Recently these data of Ruetz et al. have been retracted [*PNAS* (99) **96**, 1810], citing that *ste6* sequences were identified in yeast transformants believed to contain *PfMDR1*. Because of this report we wanted to confirm our original findings, that expression of *PfMDR1* in yeast deficient for *ste6* restores a mating phenotype, and these data are reported here. The goal of these experiments was to (1) confirm the previously observed mating phenotype; (2) demonstrate that mating is due to the presence of

PfMDR1; and (3) show that mating is not due to the presence of *ste6*. Similar experiments are independently being conducted in other laboratories to confirm these results.

Two independently derived plasmid constructs containing the *PfMDR1* gene were tested (pY*PfMDR1*-1 and pY*PfMDR1*-2). As controls, the same plasmid containing either no insert (pY) or the *ste6* gene (pY*ste6*) was used. Yeast strains used were the *ste6*-deficient (*ste6*) strain SM1563 [**a** *trp*1 *leu*2 *ura*3 *his*4 *can*1 *ste6:LEU2*) into which plasmids were transformed, and the MAT strain SM1068 [*lys*1] to test the ability of the transformants to confer a mating phenotype. Three independent mating assays were performed with three single colonies from new transformation experiments for each of the two *PfMDR1* plasmids. Mating assays were performed by mixing 10^7 MAT**a** cells with 10^8 MAT cells, and plating the mixture on SD plates. The number of diploids formed after two days was counted with the data from three mating assay shown in **Table II**. These data confirm that yeast containing both of the *PfMDR1* plasmids were able to complement the STE6 phenotype and restore the ability of yeast deficient for *ste6* to mate.

It was observed that on average, only approximately one out of every ten to fifty transformants resulted in successful complementation of mating phenotype (data not shown). The reason for this is not known, but presumably plasmid loss or rearrangement results in the loss of *PfMDR1* expression in these cells. Freshly transformed cells were more likely to yield transformants that complemented *ste6*, and for these experiments, three of five colonies isolated for each plasmid (pY*PfMDR1*-1 or pY*PfMDR1*-2) successfully mated. Experiments performed in our laboratory used yeast strains with distinct genetic backgrounds, and different plasmid constructs expressing *PfMDR1* that were derived independently from those used in the work by Ruetz et al. When plasmids containing *PfMDR1* sequences received from Dr. Phillipe Gros (pVT-*PfMDR*) were tested in our yeast assay system, these transformants did not restore a mating phenotype (**Table II**).

Table II: Summary of Mating Phenotype for Yeast Transformed with *PfMDR1*

Yeast Transformant	I	II	III	IV
pY	0	0	0	0
pY*ste6*	>2000*	>1500*	>2000*	>2000*
pY*PfMDR1*-1.1	207	149	115	79
pY*PfMDR1*-1.2	241	139	228	
pY*PfMDR1*-1.3	189	186	153	
pY*PfMDR1*-2.1	95	82	196	
pY*PfMDR1*-2.2	133	114	134	
pY*PfMDR1*-2.3	119	109	177	

Yeast Transformant	IV	V
pY	0	0
pY*ste6*	>2000*	>2000*
pY*PfMDR1*	79	101
pVT-*PfMDR*-1	0	0
pVT-*PfMDR*-2	0	0

Yeast transformed with *PfMDR1* plasmid conferred a mating phenotype. MAT**a** yeast deficient for *ste6* (SM1563) were transformed with pY*PfMDR1*, and three separate transformants for each of the two independently derived yeast expression plasmids containing *PfMDR1* (pY*PfMDR1*-1 and pY*PfMDR1*-2) were analyzed. In three separating mating assays, a total of 10^7 MAT**a** cells and 10^8 MAT cells were incubated and the number of diploids recovered for each experiment (I-III)are reported. Additional experiments using two pVT-PfMDR plasmids (Ruetz et al) were performed (IV-V). The number of diploids for the *ste6* control was estimated by plating a dilution of the mixture.

To test if the observed mating phenotype was due to the presence of the *PfMDR1* gene, and to demonstrate that the native *ste6* gene is not present in these yeast transformants, experiments using the polymerase chain reaction (PCR) were performed using gene specific primers. This analysis was performed both on total DNA derived from yeast transformants that had successfully mated, as well as on plasmid DNA recovered from bacteria transformed with a sample of this total DNA. Primer sequences for *ste6* were derived from nucleotides 755-780 and 2069-2095 and amplified a product of approximately 1340 nucleotides, while primer sequences for *PfMDR1* were derived from nucleotides 510-534 and 1462-1487 and amplified a product of approximately 980 nucleotides. These data demonstrate that DNA derived from yeast that conferred a mating phenotype contained *PfMDR1* sequences for yeast transformed with either pY*PfMDR1*-1 or pY*PfMDR1*-2, and did not contain contaminating *ste6* sequences. Similarly, plasmids derived from bacteria transformed with these DNA samples contained the expected *PfMDR1* sequences from yeast transformed with either pY*PfMDR1*-1 or pY*PfMDR1*-2, and did not contain *ste6* sequences. These data demonstrate that yeast transformed with *PfMDR1* that conferred a mating phenotype contain *PfMDR1*, but not *ste6*. Together these data demonstrate that yeast deficient for *ste6* that are transformed with *PfMDR1* restore a mating phenotype, and that this mating phenotype is due to the presence of *PfMDR1* and not contaminating *ste6* sequences.

(6) Key Research Accomplishments

Development of the Serial Analysis of Gene Expression (SAGE) system for *Plasmodium falciparum*

Analysis of 5' UTR of *Plasmodium falciparum* genes

Identification and Functional Analysis of Plasmodium cis-regulatory elements for gene expression

Development of Yeast Microarray

(7) Reportable Outcomes

Manuscripts

Golightly LM, Mbacham W, Daily J, Wirth DF. 3' UTR elements enhance expression of Pgs28, an ookinete protein of *Plasmodium gallinaceum*. Molec Biochem Parasit. 2000; 105:61-70.

Nau, ME, Lyndal R. Emerson, LR, Martin RK, Kyle DE, Wirth DF, Vahey M. Technical Assessment of the Affymetrix Yeast expression GeneChip YE6100 platform in a heterologous model of genes that confer resistance to antimalarial drugs in yeast. J Clin Microbiol. 2000; 38:1901-1908.

Laserson KF, Petralanda I, Almera R, Barker RH Jr, Spielman A, Maguire JH, Wirth DF. Genetic characterization of an epidemic of *P. falciparum* malaria among Yanomami Amerindians. J Infec Dis. 1999; 180:2081-2085.

Patankar S, Fujioka H, Wirth DF. The signal sequence and C-terminal hydrophobic domain are required for localization of the sexual stage antigen Pgs28 to the surface of *P. gallinaceum* ookinetes. Molec Biochem Parasit. 2000; 111:425-435.

Munasinghe A, Patankar S, Cook BP, Madden SL, Martin RK, Kyle DE, Cummings LM, Wirth DF. Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A-T rich genomes. Molec Biochem Parasit. 2000; in press.

Mbacham WF, Chow CS, Daily J, Golightly LM, Wirth DF. Deletion analysis of the 5' flanking sequence of the *Plasmodium gallinaceum* sexual stage specific gene *pgs28* suggests a bipartite arrangement of *cis*-control elements. Molec Biochem Parasit. 2000; in press.

Patankar S, Munasinghe A, Shoaibi A, Cummings LM, Wirth DF. Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages, as well as novel transcriptional phenomenon in the malarial parasite. 2000; submitted.

Abstracts

Patankar, S, Munasinghe A, Martin RK, Kyle D, Wirth DF. 1999. Serial analysis of gene expression in *Plasmodium falciparum*. Molecular Parasitology Meeting, Woods Hole, MA.

Chow CS, Mbacham W, Wirth DF. 1999. Analysis of the *Plasmodium gallinaceum* sexual stage specific gene *pgs28* promoter. Molecular Parasitology Meeting, Woods Hole, MA.

Thomas SM, Maguire JH, Wirth DF. 1999. Do mutations in the *MutS* gene of *Plasmodium falciparum* lead to the development of drug resistance? Molecular Parasitology Meeting, Woods Hole, MA.

Patankar S, Munasinghe A, Martin RK, Kyle D, Wirth DF. 1999. Study of transcriptional responses in *Plasmodium falciparum* by serial analysis of gene expression. American Society of Tropical Medicine and Hygiene, Washington, D.C.

Vieira PP, Alween A, Marques C, Pereira T, Wirth DF, Zalis MG. 1999. Mutation analysis of the cg2 gene in *Plasmodium falciparum* isolates from the Brazilian Amazon region. American Society of Tropical Medicine and Hygiene, Washington, D.C.

Emerson L, Nau M, Martin K, Vahey M, Kyle D, Wirth D. 1999. Global expression response elicited by anti-malarial drugs in yeast, a model system. American Society of Tropical Medicine and Hygiene, Washington, D.C.

Volkman SK, Wirth DF. 1999. Analysis of *PfMDR1* expression in a heterologous yeast system. 1999. American Society of Tropical Medicine and Hygiene, Washington, D.C.

Myrick A, Wirth DF. 2000. Analysis of the 5' upstream region of the *Plasmodium falciparum* multidrug resistance gene. Molecular Parasitology Meeting, Woods Hole, MA.

Chow S, Wirth DF. 2000. Analysis of the *Plasmodium gallinaceum* stage specific gene *pgs28* promoter. Molecular Parasitology Meeting, Woods Hole, MA.

Dodge MA, Volkman SK, Wirth DF. 2000. Subcellular localization of *Le*MDR1. Molecular Parasitology Meeting, Woods Hole, MA.

Volkman SK, Wirth DF. 2000. Structural analysis of *PfMDR1* mRNA transcripts in *Plasmodium falciparum.* Molecular Parasitology Meeting, Woods Hole, MA.

Patankar S, Munasinghe A, Martin RK, Cummings L, Wirth DF. 2000. Study of transcriptional responses to drug pressure in *Plasmodium falciparum* using SAGE. Molecular Parasitology Meeting, Woods Hole, MA.

Thomas SM, Maguire JH, Wirth DF. 2000. Are defects in the mismatch repair system of *Plasmodium falciparum* involved in the development of drug resistance? American Society of Tropical Medicine and Hygiene, Houston, TX.

Munasinghe A, Patankar S, Martin RK, Kyle D, Cummings L, Wirth DF. 2000. Investigation of chloroquine-induced alterations in global transcript profiles of *Plasmodium falciparum* using serial analysis of gene expression. American Society of Tropical Medicine and Hygiene, Houston, TX

Emerson LR, Nau ME, Martin RK, Kyle DE, Vahey M, Wirth DF. 2000. Iron transport and chloroquine toxicity. American Society of Tropical Medicine and Hygiene, Houston, TX.

Zaman M, Wirth DF. 2000. Characterization of the requirements for full drug resistance by the *Leishmania enriettii* V-circle. American Society of Tropical Medicine and Hygiene, Houston, TX.

Volkman SK, Daily JP, Sen P, Wirth DF. 2000. ABC transporters in *Plasmodium falciparum*. American Society of Tropical Medicine and Hygiene, Houston, TX.

Invited Speaking Engagements

Wirth DF, Chow L, Emerson L, Kuchler K, Volkman S. 1999. ABC transporters in leishmania and plasmodium. ATP-Binding Cassette (ABC) proteins: From multidrug resistance to genetic disease. Gosau, Austria.

Wirth DF. Invited Speaker. XVII Whitehead Symposium, Biology of Drug Discovery, October 24-26, 1999, Massachusetts Institute of Technology, Cambridge, MA

Wirth DF. Invited Speaker. Gordon Research Conference, June 20-25, 1999, Newport, R.I.

Wirth DF. Invited Speaker. Molecular Approaches to Malaria, February 2-5, 2000. Lorne, Victoria, Australia

Wirth DF. Invited Keynote Speaker. World Health Week, March 23, 2000. Vanderbilt University, Nashville, TN

Wirth DF. Invited Speaker. Malaria Genome Sequencing Meeting, June 4-6, 2000. Welcome Trust, Hinxton, Camridgeshire, U.K.

Wirth DF. Invited Speaker. Oxford 2000 Conference, September 18-22, 2000. Oxford, U.K.

Wirth DF. Invited Speaker. Infectious Disease: Challenges and New Strategies for Control – a Scientific Symposium, September 22, 2000. Harvard Medical School, Boston, MA

Wirth DF. Invited Speaker. American Society of Tropical Medicine and Hygiene, October 29 – November 2, 2000. Houston, TX

(8) CONCLUSIONS

Malaria represents a major and increasing threat to the U.S. Military. Many of the sites of current or potential U.S. Military involvement are endemic for malaria and in

several sites; multidrug resistant *P. falciparum* represents a major problem especially for non-immune military personnel. Current drugs available to the U.S. Military are quickly losing their effectiveness because of emerging and spreading drug resistance. This work is directed both at identifying new drugs and drug targets, but equally importantly toward an understanding of drug resistance mechanisms with the goal of preventing or overcoming drug resistance in the malaria parasite.

A new strategy for drug development is urgently needed. Current drugs are based on a small number of target molecules or lead compounds and in most cases the target of drug action is yet to be identified. Resistance is emerging rapidly and the mechanisms of resistance are poorly understood. The identification of new targets or new candidate drugs based on an understanding of the parasite biology are key elements in this new strategy. Clearly the development of a new antimalarial will require both basic and applied research working in concert with one another.

The goal of this work is to use a molecular genetic approach both in the identification of new drug targets and in the investigation of mechanisms of drug resistance. Progress has been made in several key areas. During this year we have tried new technical approaches to address the key goals of this work. These technical approaches were not available at the time of the original plan and are based on the rapidly evolving genome projects, including the completion of the yeast genome sequence and the development of the Plasmodium falciparum genome project. We have used these advances both in developing methods for understanding gene expression in response to drug treatment and in the future hope to use these methods to identify new drug targets.

(9) REFERENCES
Given for each section

5.1 Functional analysis of putative drug resistance genes and new drug target genes in the malaria parasite through the further development of a transformation system for the malaria parasite (pg. 6-14)

Bowman, S., D. Lawson, D. Basham, D. Brown, T. Chillingworth, C. M. Churcher, A. Craig, R. M. Davies, K. Devlin, T. Feltwell, S. Gentles, R. Gwilliam, N. Hamlin, D. Harris, S. Holroyd, T. Hornsby, P. Horrocks, K. Jagels, B. Jassal, S. Kyes, et al. (1999). "The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum [see comments]." Nature **400**(6744): 532-8.

Carulli, J. P., M. Artinger, P. M. Swain, C. D. Root, L. Chee, C. Tulig, J. Guerin, M. Osborne, G. Stein, J. Lian and P. T. Lomedico (1998). "High throughput analysis of differential gene expression." J Cell Biochem Suppl **31**: 286-96.

DeRisi, J., B. van den Hazel, P. Marc, E. Balzi, P. Brown, C. Jacq and A. Goffeau (2000). "Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants." FEBS Lett **470**(2): 156-60.

DeRisi, J. L., V. R. Iyer and P. O. Brown (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-6.

Gardner, M. J., H. Tettelin, D. J. Carucci, L. M. Cummings, L. Aravind, E. V. Koonin, S. Shallom, T. Mason, K. Yu, C. Fujii, J. Pederson, K. Shen, J. Jing, C. Aston, Z. Lai, D. C. Schwartz, M. Pertea, S. Salzberg, L. Zhou, G. G. Sutton, et al. (1998). "Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum [published erratum appears in Science 1998 Dec 4;282(5395):1827]." Science **282**(5391): 1126-32.

Hayward, R. E., J. L. Derisi, S. Alfadhli, D. C. Kaslow, P. O. Brown and P. K. Rathod (2000). "Shotgun DNA microarrays and stage-specific gene expression in Plasmodium falciparum malaria." Mol Microbiol **35**(1): 6-14.

Heller, R. A., M. Schena, A. Chai, D. Shalon, T. Bedilion, J. Gilmore, D. E. Woolley and R. W. Davis (1997). "Discovery and analysis of inflammatory disease-related genes using cDNA microarrays." Proc Natl Acad Sci U S A **94**(6): 2150-5.

Hibi, K., Q. Liu, G. A. Beaudry, S. L. Madden, W. H. Westra, S. L. Wehage, S. C. Yang, R. F. Heitmiller, A. H. Bertelsen, D. Sidransky and J. Jen (1998). "Serial analysis of gene expression in non-small cell lung cancer." Cancer Res **58**(24): 5690-4.

Hibi, K., W. H. Westra, M. Borges, S. Goodman, D. Sidransky and J. Jen (1999). "PGP9.5 as a candidate tumor marker for non-small-cell lung cancer." Am J Pathol **155**(3): 711-5.

Lal, A., A. E. Lash, S. F. Altschul, V. Velculescu, L. Zhang, R. E. McLendon, M. A. Marra, C. Prange, P. J. Morin, K. Polyak, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, R. L. Strausberg and G. J. Riggins (1999). "A public database for gene expression in human cancers." Cancer Res **59**(21): 5403-7.

Lashkari, D. A., J. L. DeRisi, J. H. McCusker, N. A. F., C. Gentile, S. Y. Hwang, P. O. Brown and R. W. Davis (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis." Proceedings of the National Academy of Sciences, USA **94**(24): 13057-13062.

Liang, P. and A. B. Pardee (1992). "Differential Display of Eukaryotic Messenger RNA by Means of Polymerase Chain Reaction." Science **257**: 967-971.

Madden, S. L., E. A. Galella, J. Zhu, A. H. Bertelsen and G. A. Beaudry (1997). "SAGE transcript profiles for p53-dependent growth regulation." Oncogene **15**(9): 1079-85.

Matsumura, H., S. Nirasawa and R. Terauchi (1999). "Technical advance: transcript profiling in rice (Oryza sativa L.) seedlings using serial analysis of gene expression (SAGE) [In Process Citation]." Plant J **20**(6): 719-26.

Picot, S., J. Burnod, V. Bracchi, B. F. Chumpitazi and P. Ambroise-Thomas (1997). "Apoptosis related to chloroquine sensitivity of the human malaria parasite Plasmodium falciparum." Trans R Soc Trop Med Hyg **91**(5): 590-1.

Polyak, K., Y. Xia, J. L. Zweier, K. W. Kinzler and B. Vogelstein (1997). "A model for p53-induced apoptosis [see comments]." Nature **389**(6648): 300-5.

Schena, M., R. A. Heller, T. P. Theriault, K. Konrad, E. Lachenmeier and R. W. Davis (1998). "Microarrays: biotechnology's discovery platform for functional genomics [see comments]." Trends Biotechnol **16**(7): 301-6.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray [see comments]." Science **270**(5235): 467-70.

Schena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown and R. W. Davis (1996). "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes." Proc Natl Acad Sci U S A **93**(20): 10614-9.

Srivastava, I. K., J. M. Morrisey, E. Darrouzet, F. Daldal and A. B. Vaidya (1999). "Resistance mutations reveal the atovaquone-binding domain of cytochrome b in malaria parasites." Mol Microbiol **33**(4): 704-11.

Srivastava, I. K. and A. B. Vaidya (1999). "A mechanism for the synergistic antimalarial action of atovaquone and proguanil." Antimicrob Agents Chemother **43**(6): 1334-9.

Thelu, J., J. Burnod, V. Bracchi and P. Ambroise-Thomas (1994). "Identification of differentially transcribed RNA and DNA helicase-related genes of Plasmodium falciparum." DNA Cell Biol **13**(11): 1109-15.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). "Serial analysis of gene expression [see comments]." Science **270**(5235): 484-7.

Velculescu, V. E., L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, P. Hieter, B. Vogelstein and K. W. Kinzler (1997). "Characterization of the Yeast Transcriptosome." Cell **88**: 243-251.

Zhang, L., W. Zhou, V. E. Velculescu, S. E. Kern, R. H. Hruban, S. R. Hamilton, B. Vogelstein and K. W. Kinzler (1997). "Gene expression Profiles in Normal and Cancer Cells." Science **276**: 1268-1272.

5.2 Analysis of Gene Expression in Plasmodium falciparum (pg. 13-24)

[1] Su X, Wellems T. Sequence, transcript characterization and polymorphisms of a *Plasmodium falciparum* gene belonging to the heat-shock protein (HSP) 90 family. Gene 1994; 151: 225-30.

[2] Horrocks P, Kilbey B. Physical and functional mapping of the transcriptional start sites of *Plasmodium falciparum* proliferating cell nuclear antigen. Mol Biochem Parasitol 1996; 82: 207-15.

[3] Horrocks P, Dechering K, Lanzer M. Control of gene expression in *Plasmodium falciparum*. Mol Biochem Parasitol 1998; 95: 171-81.

[4] Crabb BS, Cowman AF. Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*. Proc Natl Acad Sci USA 1996; 93: 7289-94.

[5] Dechering KJ, Kaan AM, Mbacham W, Wirth DF, Eling W, Konings RN, Stunnenberg HG. Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. Mol Cell Biol 1999; 19: 967-78.

[6] Horrocks P, Lanzer M. Mutational analysis identifies a five base pair cis-acting sequence essential for *GBP130* promoter activity in *Plasmodium falciparum*. Mol. Biochem. Parasitol 1999; 99: 77-87.

[7] Adini A, Warburg A. Interaction of *Plasmodium gallinaceum* ookinetes and oocysts with extracellular matrix proteins. Parasitol. 1999; 119: 331-6.

[8] Sidén-Kiamos I, Vlachou D, Margos G, Beetsma A, Waters A Sinder R, Louis C. Distinct roles for Pbs21 and Pbs25 in the *in vitro* ookinete to oocyst transformation of *Plasmodium berghei*. J Cell Sci. 2000; 113:3419-26.

[9] Matsuoka H, Kobayashi J, Barker G, Miura K, Chinzei Y, Miyajima S, Ishii A, Sinden R. Induction of anti-malarial transmission blocking immunity with a recombinant ookinete surface antigen of *Plasmodium berghei* produced in silkworm larvae using the baculovirus expression vector system. Vaccine 1996; 14: 120-6.

[10] Kaslow D, Quakyi I, Syin C, Raum M, Keister D, Coligan J, McCutchan T, Miller LH. A vaccine candidate from the sexual stage of human malaria that contains EGF-like domains. Nature 1988; 333: 74-6.

[11] Duffy P, Pimenta P, Kaslow D. Pgs28 belongs to a family of epidermal growth factor-like antigens that are targets of malaria transmission-blocking antibodies. J Exp Med 1993; 177: 505-10.

[12] Paton M, Barker G, Matsuoka H, Ramesar J, Janse C, Waters A, Sinden R. Structure and expression of a post-transcriptionally regulated malaria gene encoding a surface protein from the sexual stages of *Plasmodium berghei*. Mol Biochem Parasitol 1993; 59: 263-75.

[13] Vervenne R, Dirks R, Ramesar J, Waters A, Janse C. Differential expression in blood stages of the gene coding for the 21-kilodalton surface protein of ookinetes of *Plasmodium berghei* as detected by RNA *in situ* hybridisation. Mol Biochem Parasitol 1994; 68: 259-66.

[14] Thompson J, Sinden R. *In situ* detection of Pbs21 mRNA during sexual development of *Plasmodium berghei*. Mol Biochem Parasitol 1994; 68: 189-96.

[15] Goonewardene R, Daily J, Kaslow D, Sullivan TJ, Duffy P, Carter R, Mendis K, Wirth D. Transfection of the malaria parasite and expression of firefly luciferase. Proc Natl Acad Sci USA 1993; 90: 5234-36.

[16] Golightly L, Mbacham W, Daily J, Wirth D. 3' UTR elements enhance expression of Pgs28, an ookinete protein of *Plasmodium gallinaceum*. Mol Biochem Parasitology 2000; 105: 61-70.

[17] Patankar S, Fujioka H, Wirth D. Localization of Pgs28 to the surface of *P. gallinaceum* ookinetes requires the signal sequence and C-terminal hydrophobic domain. Mol Biochem Parasitol 2000; 111:425-35.

(5.3) Functional analysis of putative drug resistance and new drug target genes in the heterologous yeast expression system (pg. 25-30)

Delling, Raymond, Shurr (1998). *Antimicrobial Agents and Chemotherapy.* 42(5), 1034-1041.

Jelinsky and Samson (1999). *Proc. Natl. Acad. Sci.* 96, 1486-1491.

Johnston, M. (1998). "Gene chips: Array of hope for understanding gene regulation." Current Biology, 8(5), R171-R174.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., F., N. A., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis." Proceedings of the National Academy of Sciences, USA, 94(24), 13057-13062.

Lesuisse , Simon-Casteras, Labbe (1998) *Microbiology*, 144, 3455-3462. Johnston, M. (1998). "Gene chips: Array of hope for understanding gene regulation." Current Biology, 8(5), R171-R174.

Liang, P., and Pardee, A. B. (1992). "Differential Display of Eukaryotic Messenger RNA by Means of Polymerase Chain Reaction." Science, 257, 967-971.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). "Serial Analysis of Gene Expression." Science, 270, 484-487.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). "Characterization of the Yeast Transcriptosome." Cell, 88, 243-251.

(10) APPENDIX

DAMD17-98-1-8003 "New Strategies for Drug Discovery and Development for *Plasmodium falciparum*"

Principal Investigator: Dyann F. Wirth, Ph.D.

- Patankar S, Fujioka H, Wirth DF. The signal sequence and C-terminal hydrophobic domain are required for localization of the sexual stage antigen Pgs28 to the surface of *P. gallinaceum* ookinetes. Mol Biochem Parasit. 2000; 111:425-435.

- Nau ME, Emerson LR, Martin RK, Kyle DE, Wirth DF, Vahey M. Technical assessment of the affymetrix yeast expression GeneChip YE6100 platform in a heterologous model of genes that confer resistance to antimalarial drugs in yeast. J Clin Microbiol. 2000; 38:1901-1908.

- Munasinghe A;, Patankar S, Cook BP, Madden SL, Martin RK, Kyle DE, Shoaibi A, Cummings LM, Wirth DF. Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A-T rich genomes. Mol Biochem Parasit. 2001; in press.

- Mbacham WF, Chow CS, Daily J, Golightly LM, Wirth DF. Deletion analysis of the 5' flanking sequence of the *Plasmodium gallinaceum* sexual stage specific gene *pgs28* suggests a bipartite arrangement of *cis*-control elements. Mol Biochem Parasit. 2001; in press.

- Patankar S, Munasinghe A, Shoaibi, Cummings LM, Wirth DF. Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages, as well as novel transcriptional phenomenon in the malarial parasite. Submitted.

ELSEVIER

Molecular and Biochemical Parasitology 111 (2000) 425–435

MOLECULAR AND BIOCHEMICAL PARASITOLOGY

www.parasitology-online.com.

The signal sequence and C-terminal hydrophobic domain are required for localization of the sexual stage antigen Pgs28 to the surface of *P. gallinaceum* ookinetes

Swati Patankar [a], Hisashi Fujioka [b], Dyann F. Wirth [a,*]

[a] *Department of Immunology and Infectious Diseases, Harvard School of Public Health, 665 Huntington Ave., Boston, MA 02115, USA*

[b] *Institute of Pathology, Case Western Reserve University, 2085 Adelbert Road, Cleveland, OH 44106, USA*

Received 26 June 2000; received in revised form 7 September 2000; accepted 11 September 2000

Abstract

The Pgs28 protein is a major surface antigen of the sexual stages of *Plasmodium gallinaceum* — the zygotes and the ookinetes. The protein contains conserved motifs, namely an N-terminal signal sequence, four epidermal growth factor-like repeats and a C-terminal hydrophobic domain that serves as a signal for glycosylphosphatidylinositol (GPI) — anchor modification. In this study, we define the protein motifs required for the surface localization of Pgs28 in ookinetes, using transient transfection combined with immunofluorescence and confocal microscopy. Pgs28 fused to the green fluorescent protein (Pgs28-GFP) is expressed in zygotes, intermediate retort forms and ookinetes. Mutational analyses of Pgs28 coding regions reveal that deletions of the signal sequence and the C-terminal domain result in intracellular retention of the fusion protein. Therefore, the signal sequence and C-terminal domain are required for cell surface localization. Additionally, the Pgs28-GFP fusion proteins are shed from the surface of live ookinetes, suggesting that Pgs28 may be involved in interactions with the cells of the mosquito midgut or during motility.

Keywords: *Plasmodium gallinaceum*; Sexual stages; Pgs28; Signal sequence; GPI — anchor; Membrane shedding

1. Introduction

The Pxs21/25 proteins (21–25 kDa proteins found in all *Plasmodium* species) are targets for transmission-blocking vaccines [1]. These proteins include the Pfs25 and Pfs28 proteins of *Plasmodium falciparum* [1], Pbs21 from *Plasmodium berghei* [2,3] and Pgs28 from *Plasmodium gallinaceum* [4]. Numerous studies have focused on the transmission-blocking abilities of antibodies

Abbreviations: EGF, epidermal growth factor; ER, endoplasmic reticulum; FITC, fluorescein isothiocyanate; GFP, green fluorescent protein; GPI, glycosylphosphatidylinositol; IFA, immunofluorescence assay; PBS, phosphate-buffered saline; Pgs28, *Plasmodium gallinaceum* sexual stage protein of 28 kDa; Pxs21/25, family of proteins expressed in the sexual stages of all *Plasmodium* species, molecular weight 21–25 kDa.

* Corresponding author. Tel.: +1-617-4321621; fax: +1-617-4324766.

E-mail address: dfwirth@hsph.harvard.edu (D.F. Wirth).

0166-6851/00/$ - see front matter
PII: S0166-6851(00)00340-6

raised against recombinant Pfs25 and Pfs28 expressed in heterologous systems [5], while protein trafficking of Pbs21 has been elegantly studied in insect cells [6]. We are interested in defining the protein motifs that are essential for targeting Pgs28 to the surface of cells that normally express the protein, *P. gallinaceum* ookinetes. These cells provide a novel biological system for the study of many proteins currently being developed as vaccine candidates. Additionally, *P. gallinaceum* is an ideal system for the analysis of Pgs28 localization as sexual stages can be readily isolated and transfected with expression vectors. Hence, rather than using a heterologous system, we have studied Pgs28 protein trafficking in the sexual stages of *P. gallinaceum*.

The Pxs21/25 proteins provide a useful tool to study trafficking in ookinetes. These proteins contain distinct motifs, including a 21 amino acid N-terminal signal sequence that is cleaved from the mature protein, four to six EGF-like repeats and a C-terminal hydrophobic region that provides the signal for anchoring to the membrane via a glycosylphosphatidylinositol (GPI) modification [4]. The C-terminal domain is cleaved prior to addition of the GPI moiety. In eukaryotic systems, signal sequences have been shown to be essential for targeting proteins to the endoplasmic reticulum (ER) [7] while EGF repeats, first described in epidermal growth factor, are important for protein–protein interactions in cell adhesion and signaling during neurological development and coagulation [8]. The importance of the signal sequence and C-terminal region of Pbs21 has been demonstrated in insect cells [6]. Deletion of the signal sequence prevented transport of Pbs21 to the ER of the insect cells while deletion of the GPI-anchor disrupted Pbs21 translocation through the ER and distribution on the cell surface. Moreover, deletion of the GPI-anchor resulted in the secretion of recombinant protein into the culture medium.

While studying protein localization to parasite membranes, it is important to appreciate that the cell surface is a dynamic structure. Shahabuddin et al. have shown that the ookinete surface is efficiently labeled with a lipophilic dye, PKH26; this dye is shed from the motile parasite as evidenced by trails behind the ookinete suggesting that the ookinete surface membrane is sloughed off during movement [9]. Although highly likely, it is unclear whether membrane-bound proteins like Pgs28 are also shed from the surface of the ookinete.

In this work, we study localization of the Pgs28 protein in the sexual stages of the chicken malarial parasite, *P. gallinaceum* and reveal three aspects of Pgs28 protein targeting. First, by deletion analysis, we show that the signal sequence and C-terminal hydrophobic region of Pgs28 are essential for cell surface localization of the protein in ookinetes. Deletion of the signal sequence (amino acids 1–21) results in cytoplasmic localization of Pgs28, suggesting that this motif directs nascent Pgs28 into the ER of the ookinete. Similarly, deletion of the C-terminal domain (amino acids 194–212) leads to Pgs28 accumulation within the ookinete. This is consistent with the requirement for a C-terminal GPI-anchor for membrane localization. Thus, protein motifs required for surface localization of Pgs28 in the native system (*P. gallinaceum*) are similar to those defined for Pbs21 in Sf9 cells, revealing universal themes in trafficking of Pxs21/25 proteins. Second, vesicular structures, shown to contain Pgs28 protein, are visualized by immunoelectron microscopy. These may be components of the ookinete trafficking machinery. Finally, we show that a Pgs28-GFP fusion protein is shed from the surface membranes of live ookinetes. These data have implications for the role of Pgs28 in parasite motility and interactions with the mosquito midgut.

2. Materials and methods

2.1. Construction of deletions in the Pgs28 coding region

The parent plasmid from which all constructs were derived was the Pgs28.1-luc vector that has been previously described [10]. Briefly, Pgs28.1-luc contains a *pgs28-luciferase* gene fusion flanked by ~1.9 kilobases of *pgs28* 5′ sequences and ~0.6 kilobases of *pgs28* 3′ sequences cloned into the

PGEM plasmid. The *pgs28-luciferase* fusion gene under the control of these flanking sequences is efficiently expressed in the sexual stages of *P. gallinaceum* [11].

All restriction enzymes were obtained from New England Biolabs (NEB) and used with the supplied buffers. Deletions in the Pgs28 coding region that preserved the reading frame of the protein were generated using a PCR-based mutagenesis strategy. Initially, the expression plasmid, Pgs28.1-luc, was digested with *Bam*HI to release the luciferase gene. Next, depending on the deletion, the *Bam*HI-digested plasmid was further incubated with *Bgl*II (deletions of the first EGF-like repeat and the signal sequence), *Sac*I (deletion of the second EGF-like repeat) or *Bsg*I (deletion of C-terminal domain). PCR reactions, using the Pgs28.1-luc plasmid as template and the following sets of primers, were performed.

2.1.1. Deletion of the first EGF-like repeat (amino acids 28–71)

SP5 — 5′ GGT TTG TGG ACA ATG G 3′
SP7 — 5′ GTG GGA TCC GAA GGT TCA TCA TCT GAA GG 3′

The *Bam*HI site is underlined. The PCR product was digested with *Bgl*II and *Bam*HI and cloned into the digested Pgs28.1-luc vector.

2.1.2. Truncation of the second EGF-like repeat (amino acids 72–103)

SP8 — 5′ CGT AGG ATC CAA AGG AAT GTG GAG AAG G 3′
−40 Universal primer — 5′ GTT TTC CCA GTC ACG ACG TTG TA 3′

The *Bam*HI cloning site is underlined. The PCR product was digested with *Sac*I and *Bam*HI and cloned into the Pgs28.1-luc vector.

2.1.3. Deletion of the C-terminal domain (amino acids 194–212)

SP9 — 5′ CTC ATA AAG GCC AAG AAG GG 3′
SP10 — 5′ CAT TGA AAA GGG ATT AGG TGC TAT TAC CTG CAC TAG GAG GTG G 3′

Nucleotides in bold and underlined differ from to the *pgs28* gene sequence; the T was included to incorporate a stop codon while the CTGCAC sequence is the binding site for the typeIIS restriction enzyme *Bsg*I. The inclusion of the *Bsg*I site alters a serine residue to alanine. The PCR product was digested with *Bam*HI and *Bsg*I and cloned into the digested Pgs28.1-luc vector.

2.1.4. Deletion of the signal sequence (amino acids 1–21)

Two separate PCR reactions were performed to generate the signal sequence deletion. The first PCR reaction used the following primers,

SP5 — described above
SP 13.1 — 5′ GGC CGG CCG GCC ATG GAC TAG GAA TTT TCA TTT TTT TAA ATA AAT G 3′

Nucleotides in bold and underlined are different from *pgs28* gene sequences; CCATGG is a *Nco*I restriction site while CAT is a start codon (antisense strand). This PCR product was digested with *Bgl*II and *Nco*I.

SP14 — 5′ ATC GTA CCA TGG GCT CCT TCA GAT GAT G 3′
Luc.seq — 5′ TCT AGA GGA TAG AAT GGC GC 3′

The *Nco*I site is in bold and underlined. This PCR product was digested with *Nco*I and *Bam*HI. The two PCR products were cloned simultaneously into *Bgl*II/*Bam*HI digested Pgs28.1-luc in a three-way ligation. The *Nco*I site introduces two additional amino acid residues (glutamine and tryptophan).

The luciferase gene was re-introduced into the deletion plasmids via the *Bam*HI site. All deletion plasmids were sequenced through the junctions as well as the coding regions that had been PCR-amplified to ensure lack of PCR-generated mutations.

To replace luciferase with green fluorescent protein (GFP), the Pgs28.1-luc plasmid and its variants were digested with *Bam*HI to drop out the luciferase gene. The GFP (Superglow GFP [12]) coding region was amplified with the follow-

ing primers from the pJH23 yeast expression vector. *Bgl*II sites are underlined.

SP 11 — 5′ GGG ATA GGG AGA TCT AAT GGC TAG CAA AGG AG 3′

SP 12 — 5′ GGG TTT AGA TCT AAG CAG CCG GAT CCT TTG TC 3′

The resulting PCR product was digested with *Bgl*II and ligated into the bacterial expression vector, pRSET C (Invitrogen) at the unique *Bgl*II site. Expression of GFP was confirmed by transforming JM109(DE3) *Escherichia coli* (Promega) with the pRSET-GFP construct and plating on LB medium with 100 μg ml^{-1} ampicillin. The presence of low levels of T7 polymerase in JM109 (DE3) cells results in expression of GFP from the pRSET vector and this GFP expression was assessed by placing the bacterial plate on a low wavelength UV trans-illuminator. GFP-positive colonies emitted a green fluorescence when exposed to UV light. The pRSET-GFP plasmid from these GFP positive colonies was isolated and digested with *Bgl*II to release the GFP coding region; due to compatibility of *Bgl*II and *Bam*HI overhangs, GFP was then ligated into *Bam*HI-digested Pgs28.1 plasmid and the deletion vectors.

The coding regions of all Pgs28 deletion plasmids were sequenced and no PCR-generated errors detected.

2.2. Parasites and transfections

P. gallinaceum parasites were propagated in White leghorn chickens by serial injection into wing veins. At parasitemias of 50–70%, blood was withdrawn by heart puncture. Gametogenesis was induced as described previously [10], with the inclusion of xanthurenic acid (Sigma) [13] at a final concentration of 50 μM in the exflagellation buffer. Gametes and zygotes were purified, also as described previously, and 1×10^7 cells were electroporated (BioRad) with 100 μg of DNA (QIAGEN) at settings of 25 μF and 0.5 kV in 0.2 cm cuvettes (BioRad). Parasites were incubated at 25°C in Medium 199 (Gibco-BRL) and harvested for analysis at approximately 48 h after transfection.

2.3. Immunofluorescence and confocal microscopy

Transfected parasites were washed once with phosphate-buffered saline (PBS) and allowed to adhere onto poly-L-lysine coated slides. After fixation in 4% paraformaldehyde, cells were either permeabilized with 0.1% Triton X-100 and 50 mM glycine in PBS for 15 min or allowed to remain non-permeabilized. Subsequently, cells were stained with primary antibodies mAbIID2-B3B3 (1:50 dilution) to recognize endogenous Pgs28 and polyclonal affinity-purified anti-GFP antibody (1:500 dilution). Secondary antibodies (goat anti-mouse-rhodamine and goat anti-rabbit fluorescein, Boehringer–Mannheim) were used at 1:250 dilutions. Confocal microscopy was performed on a BioRad MRC-1024 laser scanning confocal microscope.

2.4. Double labeling immunoelectron microscopy

Ookinetes were fixed for 30 min at 4°C with 1% formaldehyde, 0.1% glutaraldehyde in 0.1 M phosphate buffer, pH 7.4. Fixed samples were washed, dehydrated and embedded in LR White resin (Polysciences, Inc., Warrington, PA). Thin sections on nickel grids (without a supporting film) were blocked in PBSB-Tween for 30 min. The composition of PBSB-Tween is as follows, PBS was supplemented with 1% w/v bovine serum albumin fraction V and 0.01% v/v Tween 20. Labeling for the first antigen was done on face 'A' of the grid and for the second antigen on face 'B' of the grid. Briefly, grids (face A) were incubated with anti-luciferase antibody diluted 1:20 in PBSB-Tween for 2 h at 25°C. Negative controls included normal rabbit serum and PBSB-Tween applied as the primary antibody. After washing, grids were incubated at room temperature for 1 h in 10 nm gold-conjugated goat anti-rabbit IgG (Amersham Life Sciences, Arlington, IL) diluted 1:20 in PBSB-Tween, rinsed with PBSB-Tween. Grids (face B) were then incubated with anti-Pgs28 antibody diluted 1:5–1:10 in PBSB-Tween for 2 h at 25°C, and then incubated in 15 nm gold-conjugated goat anti-mouse IgG (Amersham). Negative controls included normal mouse serum and PBSB-Tween applied as the primary

antibody. Grids were then fixed with 2.5% glutaraldehyde to stabilize the gold particles. Samples were stained with uranyl acetate and lead citrate, and then examined with a Zeiss CEM902 electron microscope (Zeiss, Oberkochen, Germany).

3. Results

3.1. Pgs28 mutant proteins are expressed in sexual stage parasites

To identify the protein motifs that direct Pgs28 to the cell surface, deletion constructs were generated in expression vectors that express Pgs28 fused to different reporter genes (luciferase or GFP). Deletions were made in motifs that were predicted to be important in surface localization based on other studies, the signal sequence (ΔSS) and the C-terminal hydrophobic domain (ΔC-term). Additionally, motifs that would not be expected to play roles in localization were also analyzed as controls. Thus a deletion was generated in the first EGF-like repeat (ΔEGF1), while a truncation was made in the second EGF-like repeat (ΔEGF2). Fig. 1 shows a schematic diagram of the deletion constructs. In order to assess the levels of expression of the mutant proteins, a luciferase reporter was fused in-frame to the mutant genes via a unique *Bam*HI site in the second EGF-like repeat. Luciferase assays of sexual stage parasites transiently transfected with the deletion constructs showed that all proteins were expressed in sexual stage parasites (data not shown).

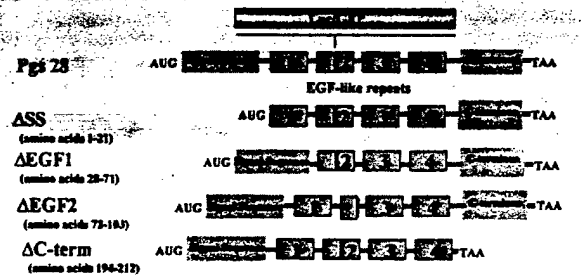


Fig. 1. Schematic of in-frame deletions of the Pgs28 protein. All proteins were expressed from the Pgs28.1 expression vector described in Goonewardene et al. AUG represents the start codon and TAA the stop codon. A unique BamHI restriction site within the second EGF-like repeat was used to clone the reporter genes into Pgs28. Reporters used were Luc, luciferase; GFP, green fluorescent protein.

The sub-cellular localization of Pgs28 deletion proteins was assessed using GFP as a reporter. Parasites were transfected with a Pgs28-GFP expression vector; localization of fusion proteins in non-permeabilized cells was assessed using immunofluorescence assays (IFA) and confocal microscopy. Double labeling of transfected parasites for both endogenous Pgs28 protein as well as the Pgs28-GFP fusion protein was performed, using secondary antibodies conjugated to rhodamine (endogenous Pgs28) and fluorescein (Pgs28-GFP) to distinguish the two signals.

Fig. 2 shows that the Pgs28-GFP fusion protein was expressed in zygotes (Panel A), ookinetes (Panel C) and the retort forms (Panel B). Approximately 1–10% of the cells showed GFP-positive staining, indicating a robust efficiency of transfection and expression. Panel A shows a zygote expressing Pgs28-GFP protein that co-localizes with endogenous Pgs28 protein; also visible is a zygote that was not transfected (arrow). In non-permeabilized ookinetes (Panel C), both Pgs28 and Pgs28-GFP show a pattern of staining that is most intense on the cell surface. Hence, Pgs28 fused to the GFP reporter shows colocalization with endogenous Pgs28, and in ookinetes that have not been permeabilized, both proteins are found predominantly on the cell surface, as expected.

Having shown that Pgs28 and Pgs28-GFP proteins localize predominantly to the cell surface, we employed a well-characterized strategy to obtain initial data regarding the localization of the mutant forms of Pgs28-GFP. Cells were processed for IFA and confocal microscopy without prior permeabilization, hence only cell surface-exposed proteins should be available for antibody recognition.

The ΔEGF1-GFP and ΔEGF2-GFP fusion proteins were present on the cell surface of non-permeabilized ookinetes, in a pattern similar to that observed for Pgs28-GFP and at a frequency similar to that observed for Pgs28-GFP (GFP-positive cells were 1–10% of total cells). Hence,

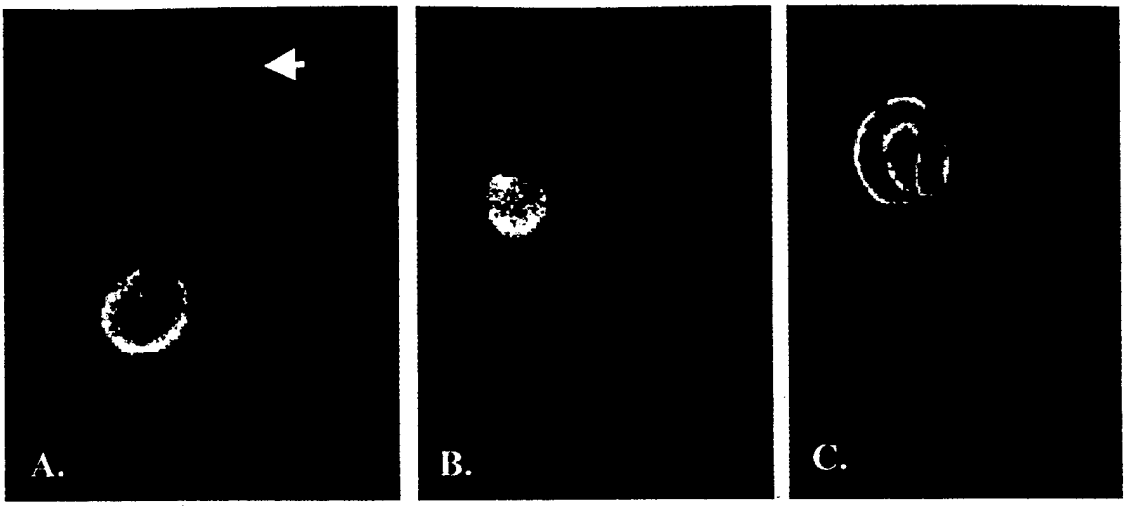


Fig. 2

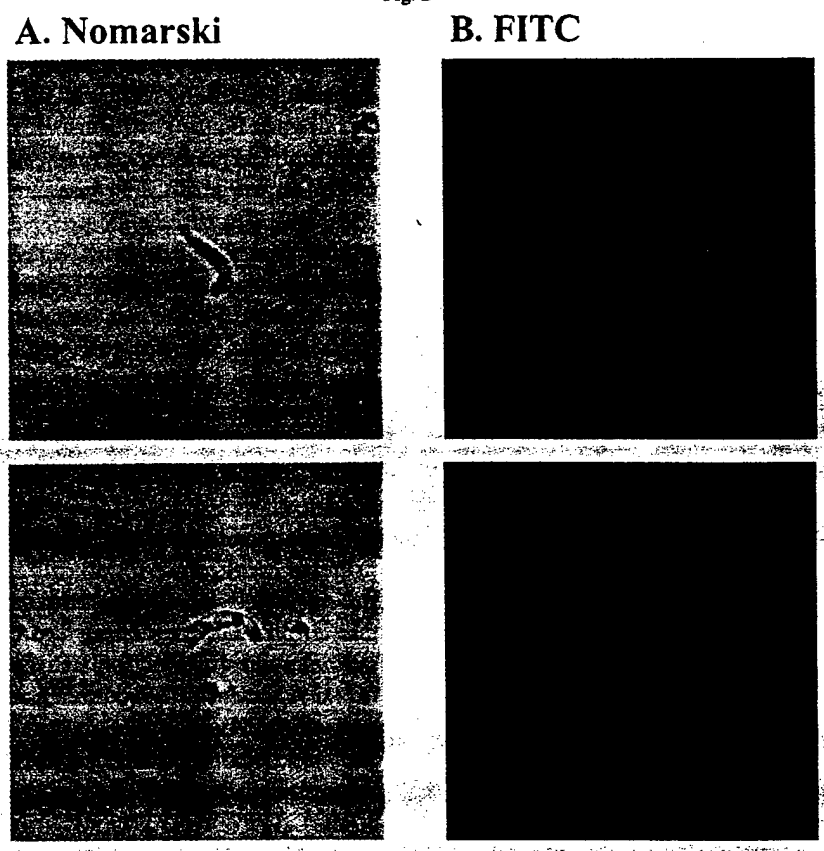


Fig. 5.

Fig. 2. IFA and confocal microscopy of sexual stages of *P. gallinaceum* expressing Pgs28-GFP fusion proteins. These cells have not been permeabilized. Panel A, zygotes; Panel B, retort form; Panel C, ookinete. Parasites were stained with antibodies against endogenous Pgs28 (detected with secondary antibodies conjugated to rhodamine, red staining) and GFP (detected with secondary antibodies conjugated to fluorescein, green staining). Colocalization of endogenous Pgs28 and Pgs28-GFP results in yellow fluorescence. The arrow denotes a zygote that has not been transfected.

Fig. 5. Live, motile ookinetes shed Pgs28-GFP from their cell surface. Ookinetes observed under Nomarski optics (Panel A), FITC filter (Panel B). The FITC images have been over exposed to reveal the fluorescent trails shed from the parasites.

deletion of the EGF-repeats appears to have little effect on localization of Pgs28. However, without permeabilization, cells transfected with an expression vector containing the ΔC-term-GFP protein exhibited no GFP fluorescence as detected by IFA and confocal microscopy. Similarly, cells transfected with an expression vector containing the ΔSS-GFP open reading frame also showed no GFP signal (GFP-positive cells not detected over hundreds of fields). From these results in non-permeabilized ookinetes, the Pgs28-GFP, ΔEGF1-GFP and ΔEGF2-GFP proteins appear to be on the surface of ookinetes while ΔC-term-GFP and ΔSS-GFP proteins may be intracellular.

3.2. The signal sequence and C-terminal hydrophobic region of Pgs28 are required for cell surface localization

To further analyze the sub-cellular localization of the Pgs28-GFP protein and its mutant derivatives, confocal microscopy was performed on permeabilized cells double-labeled with antibodies against endogenous Pgs28 and GFP. Fig. 3 shows representative images of stained ookinetes permeabilized with Triton X-100. In contrast to non-permeabilized cells shown in Fig. 2, most permeabilized cells showed intracellular staining of both endogenous Pgs28 (Fig. 3, Panel A) and the GFP fusion proteins (Fig. 3, Panels B–D). Some permeabilized ookinetes showed predominant cell surface staining (cells denoted by arrows in Fig. 3, Panels C and F). The intracellular staining may indicate the presence of Pgs28 in vesicles that comprise the trafficking machinery. The Pgs28-GFP, ΔEGF1-GFP and ΔEGF2-GFP fusion proteins showed colocalization with endogenous Pgs28 as evidenced by yellow staining on both the cell surface and internal structures (Fig. 3, Panels B–D). Moreover, endogenous Pgs28 and GFP-fusion proteins appear to be excluded from the nuclear region of ookinetes.

In contrast, deletion of the signal sequence (Fig. 3, Panel F) resulted in intracellular retention of the GFP fusion protein while endogenous Pgs28 was still localized to the surface. Staining for the ΔSS-GFP protein was diffuse, suggesting cytoplasmic localization. Deletion of the C-terminal domain also resulted in intracellular localization of GFP-fusion protein (Fig. 3, Panel E). However, in contrast to the uniformly diffuse intracellular signal obtained with the ΔSS-GFP proteins, the ΔC-term-GFP proteins appeared to be distributed within the ookinete in a similar pattern as seen for Pgs28-GFP (Fig. 3, Panel B). Hence, deletion of the both the signal sequence and the C-terminal domain results in intracellular retention of Pgs28, with ΔSS-GFP staining appearing diffuse and cytoplasmic. Similar results were obtained during IFA and confocal analysis of ΔC-term and ΔSS proteins fused to luciferase (Fig. 3, Panels G and H), as well as upon direct visualization of live parasites expressing the ΔC-term-GFP and ΔSS-GFP fusion proteins (Fig. 3, Panels I and J).

3.3. Immunoelectron microscopy defines organelles involved in trafficking of Pgs28

Immunoelectron microscopy was used to identify organellar structures through which trafficking of the Pgs28 protein occurs (Fig. 4). Analysis of ookinetes revealed that endogenous Pgs28 proteins were present on both the cell surface as well as in intracellular vesicles (Fig. 4, large gold particles) and Pgs28-luciferase fusion proteins were localized to the same compartments as endogenous Pgs28 (Fig. 4, small gold particles). Therefore, these vesicular organelles appear to be components of the trafficking machinery that transports Pgs28 to the surface membrane of the ookinete, and may constitute the ER-Golgi network.

3.4. Pgs28-GFP proteins are shed from the surface of motile ookinetes

Previous work by Shahabuddin et al. has shown that *P. gallinaceum* ookinetes can be stained with the lipophilic dye PKH26. Three to four hours post-staining, the dye begins to shed from the posterior end of the ookinetes suggesting that the ookinete surface membrane is a dynamic structure possibly due to the motile nature of the ookinete [9]. This observation suggested that localization of Pgs28 also might be affected by

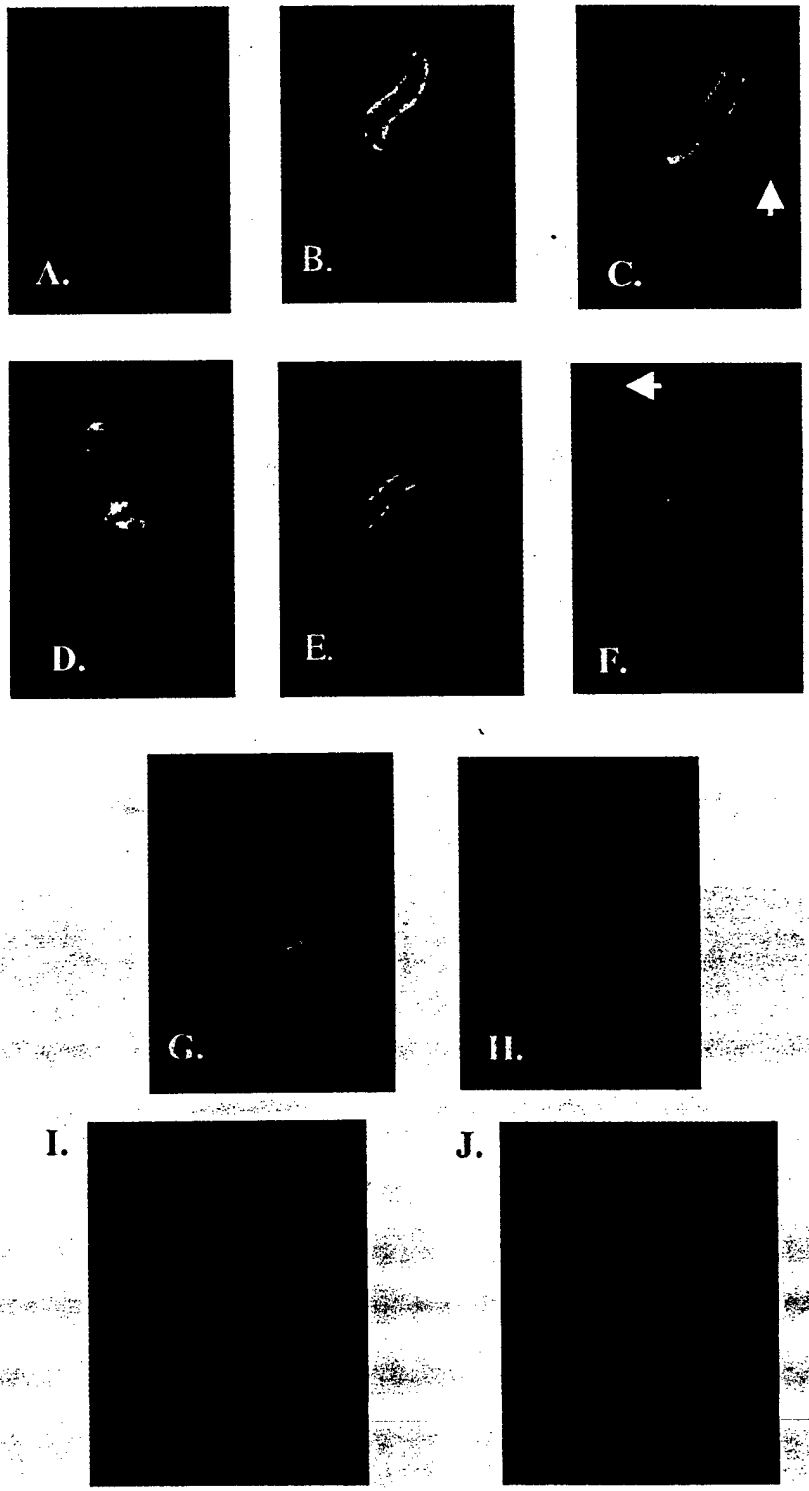


Fig. 3.

Fig. 4. Immuno-electron microscopy showing localization of endogenous Pgs28 and Pgs28-luciferase fusion proteins. Large gold particles (15 nm) label Pgs28 while small gold particles (10 nm) label Pgs28-Luc. Arrows indicate proteins present on the cell surface and internal vesicles.

motility of the ookinete and prompted us to answer whether Pgs28 was shed from the surface of migrating ookinetes. Pgs28-GFP provides a facile tool to ask this question as live parasites can be visually monitored by microscopy for localization of the GFP fusion protein. Live parasites were placed on slides for 4 h, then observed with Nomarski optics (Fig. 5, Panel A) and under a FITC filter (Fig. 5, Panel B). While no trail was distinctly visible with Nomarski optics, Pgs28-GFP fluorescence revealed that the motile ookinete shed a fluorescent trail during its movements on the slide. Similar results were obtained for both the ΔEGF1-GFP and ΔEGF2-GFP proteins but no fluorescent trails were detected when ookinetes expressed ΔSS-GFP and ΔC-term-GFP (Fig. 3, Panels I and J). These data indicate that Pgs28 is shed from the surface of motile ookinetes and are supportive of the hypothesis that deletion of the signal sequence or C-terminal domain of Pgs28 results in intracellular retention of the protein.

4. Discussion

The sexual stages of *P. gallinaceum* have been exploited effectively in previous studies on transcriptional regulation of the *pgs28* gene [11,14]. In this paper, we study trafficking of the *Pgs28* protein and reveal that cell surface localization of Pgs28 on ookinetes is critically dependent on two protein motifs, the signal sequence and the C-terminal hydrophobic domain. As might be expected, the EGF-repeats are not involved in directing localization of Pgs28.

The N-terminal signal sequence is necessary in the early stages of the trafficking pathway to the cell surface, for the targeting of nascent proteins into the ER-Golgi network. Immunofluorescence analysis shows that, consistent with this role, deletion of the signal sequence results in diffuse, cytoplasmic localization of a Pgs28-GFP fusion protein. Elegant work in *Plasmodium falciparum* has identified a bipartite signal sequence for the localization of proteins to a specialized organelle, the apicoplast [15]. The classical signal sequence of the bipartite motif is required for import into the parasite secretory pathway, while the plant-like transit peptide specifically targets proteins to the apicoplast. Similar to our data with the Pgs28 signal sequence, deletion of the signal peptide of the acyl carrier protein (ACP) results in cytoplasmic localization of GFP-tagged ACP [15].

Deletion of the C-terminal domain results in retention of Pgs28-GFP fusion proteins in internal structures, as these proteins lack the signal for insertion into the surface membrane (GPI-anchor). These internal structures have been visualized by immunoelectron microscopy (Fig. 4) and shown to contain Pgs28; these vesicles may be components of the ER/Golgi of *P. gallinaceum* ookinetes. Previous studies that performed electron microscopy on ookinetes and zygotes [16]

Fig. 3. IFA and confocal microscopy of permeabilized cells. Panel A, Cells that were mock transfected without DNA; Panel B, Pgs28-GFP; Panel C, ΔEGF1-GFP; Panel D, ΔEGF2-GFP; Panel E, ΔC-terminus-GFP; Panel F, Δ signal sequence-GFP. Panel G, ΔC-terminus-luciferase; Panel H, ΔSS-luciferase. Red staining indicates endogenous Pgs28 protein and green staining indicates GFP and luciferase fusion proteins. Arrows indicate ookinetes that express Pgs28 predominantly on the cell surface. Panel I, a live ookinete expressing ΔC-terminus-GFP; observed under FITC filter without immunostaining. Panel J, a live ookinete expressing ΔSS-GFP; also observed under FITC filter without immunostaining.

have identified vesicular structures corresponding to the ER; here we show that, indeed, vesicular structures within ookinetes do contain Pgs28 proteins. In our IFA experiments, all ookinetes show cell surface localization of Pgs28 while a subset of ookinetes does not exhibit intracellular localization of endogenous Pgs28 (Fig. 3, Panels C and F). This could be due to the fact that trafficking of Pgs28 is dynamic and fixation during immunofluorescence captures ookinetes at different stages of this process.

The requirement for the signal sequence and C-terminal domain is similar to results obtained for Pbs21 localization in a heterologous expression system [6]. Similarly, deletion of the GPI-anchor addition signal of variant surface glycoproteins (VSG) in Trypanosomes resulted in delayed forward transport of the mutant proteins and retention in the ER [17]. Taken together, these data indicate that the C-terminal domain of Pgs28 contains a signal, presumably the GPI-anchor itself, which allows efficient passage of the protein through the cellular trafficking machinery.

The results described above indicate that trafficking of at least one GPI-anchored protein (Pgs28) in ookinetes follows previously described paradigms of protein trafficking in parasitic protozoa and other eukaryotes. These common themes in protein trafficking are underscored by the fact that identical protein motifs were defined for localization of two Pxs21/25 proteins to the cell surface in two completely different systems, Pbs21 in Sf9 insect cells [6] and Pgs28 in *P. gallinaceum* ookinetes (this report). Hence, contrary to experiments in *P. falciparum* where the promoter of the AMA-1 gene played an important role in appropriate localization of the protein [18], the precise timing of expression of Pgs28 and Pbs21 is not critical to localization.

Some cell-specific characteristics do emerge while studying protein localization in ookinetes. For example, Pgs28 is shed from the surface of motile ookinetes most likely in trails that have been previously identified [9]. This result raises interesting questions regarding a potential function for Pgs28 in motility or cell–cell interactions between ookinetes and the mosquito mid-gut membranes. Based upon the functions of EGF-repeats in other systems [8], the EGF-repeats in Pgs28 will certainly be involved in protein interactions and signaling between ookinetes and midgut cells. Unpublished knockout experiments of the *pbs21* and *pbs25* genes suggest that these members of the Pxs21/25 family play limited roles in ookinete motility and invasion but may be involved in oocyst development (Andy Waters, personal communication). Similar shedding of GPI-anchored VSGs from the surface of Trypanosomes has been reported. The functional significance of this shedding is thought to be evasion of complement-mediated lysis [19] or replacement of the VSGs on the surface of *T. brucei* with procyclin proteins [20]. In *P. gallinaceum*, shedding of the membrane along with proteins like Pgs28 may also be required for immune-evasion or replacement with newly synthesized cell surface proteins.

In conclusion, this work sheds light on the protein signals required for the transport of Pgs28 to the surface of ookinetes. The data also raise questions regarding the function of Pgs28 in the development of *P. gallinaceum* within its mosquito vector. High efficiencies of transfection, ready availability of expression vectors and reporter genes, and applicability of molecular and cell biology techniques will make *P. gallinaceum* zygotes and ookinetes an excellent system for further analysis of the biology of the *pgs28* gene and other genes expressed in *Plasmodium* sexual stages.

Acknowledgements

We acknowledge Gilberto Ramirez for his expert assistance in maintaining *P. gallinaceum* strains in chickens. We thank Dr David Kaslow for providing monoclonal antibodies against Pgs28 and Drs Paul Ferrigno and Pam Silver for providing antibodies against GFP and the 'Superglow GFP' expression vector used to make the GFP fusion constructs. We also thank Drs Connie Chow and Sorab Dalal for critical reading of the manuscript. This work was funded by grants from the National Institutes of Health and Department of Defense to Dyann F. Wirth.

References

[1] Duffy PE, Kaslow DC. A novel malaria protein, Pfs28, and Pfs25 are genetically linked and synergistic as falciparum malaria transmission-blocking vaccines. Infect Immun 1997;65:1109–13.

[2] Margos G, van Dijk MR, Ramesar J, Janse CJ, Waters AP, Sinden RE. Transgenic expression of a mosquito-stage malarial protein, Pbs21, in blood stages of transformed *Plasmodium berghei* and induction of an immune response upon infection. Infect Immun 1998;66:3884–91.

[3] Spano F, Matsuoka H, Ozawa R, Chinzei Y, Sinden RE. Epitope mapping on the ookinete surface antigen Pbs21 of *Plasmodium berghei*: identification of the site of binding of transmission-blocking monoclonal antibody 13.1. Parassitologia 1996;38:559–63.

[4] Duffy PE, Pimenta P, Kaslow DC. Pgs28 belongs to a family of epidermal growth factor-like antigens that are targets of malaria transmission-blocking antibodies. J Exp Med 1993;177:505–10.

[5] Gozar MM, Price VL, Kaslow DC. Saccharomyces cerevisiae-secreted fusion proteins Pfs25 and Pfs28 elicit potent *Plasmodium falciparum* transmission-blocking antibodies in mice. Infect Immun 1998;66:59–64.

[6] Blanco AR, Paez A, Gerold P, Dearsly AL, Margos G, Schwarz RT, Barker G, Rodriguez MC, Sinden RE. The biosynthesis and post-translational modification of Pbs21 an ookinete-surface protein of *Plasmodium berghei*. Mol Biochem Parasitol 1999;98:163–73.

[7] Wilkinson BM, Regnacq M, Stirling CJ. Protein translocation across the membrane of the endoplasmic reticulum. J Membr Biol 1997;155:189–97.

[8] Davis CG. The many faces of epidermal growth factor repeats. New Biol 1990;2:410–9.

[9] Shahabuddin M, Gayle M, Zieler H, Laughinghouse A. *Plasmodium gallinaceum*: fluorescent staining of zygotes and ookinetes to study malaria parasites in mosquito. Exp Parasitol 1998;88:79–84.

[10] Goonewardene R, Daily J, Kaslow D, Sullivan TJ, Duffy P, Carter R, Mendis K, Wirth D. Transfection of the malaria parasite and expression of firefly luciferase. Proc Natl Acad Sci USA 1993;90:5234–6.

[11] Golightly LM, Mbacham W, Daily J, Wirth DF. 3' UTR elements enhance expression of Pgs28, an ookinete protein of *Plasmodium gallinaceum*. Mol Biochem Parasitol 2000;105:61–70.

[12] Kahana JA, Silver PA. Use of *A. victoria* green fluorescent protein to study protein dynamics in vivo. In: Ausubel FM, et al., editors. Current Protocols in Molocular Biology, vol. 1. New York: Wiley, 1996:9.6.13–19.16.19.

[13] Billker O, Lindo V, Panico M, Etienne AE, Paxton T, Dell A, Rogers M, Sinden RE, Morris HR. Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. Nature 1998;392:289–92.

[14] Dechering KJ, Kaan AM, Mbacham W, Wirth DF, Eling W, Konings RN, Stunnenberg HG. Isolation and functional characterization of two distinct sexual-stage- specific promoters of the human malaria parasite *Plasmodium falciparum*. Mol Cell Biol 1999;19:967–78.

[15] Waller RF, Reed MB, Cowman AF, McFadden GI. Protein trafficking to the plastid of plasmodium falciparum is via the secretory pathway. EMBO J 2000;19:1794–802.

[16] Sinden RE. Gametocytes and sexual development. In: Sherman IW, editor. Malaria: Parasite Biology, Pathogenesis and Protection. Cleveland, OH: American Society of Metals Press, 1998:25–48.

[17] McDowell MA, Ransom DM, Bangs JD. Glycosylphosphatidylinositol-dependent secretory transport in *Trypanosoma brucei*. Biochem J 1998;335:681–9.

[18] Kocken CH, van der Wel AM, Dubbeld MA, Narum DL, van de Rijke FM, van Gemert GJ, van der Linde X, Bannister LH, Janse C, Waters AP, Thomas AW. Precise timing of expression of a *Plasmodium falciparum*-derived transgene in *Plasmodium berghei* is a critical determinant of subsequent subcellular localization. J Biol Chem 1998;273:15119–24.

[19] Frevert U, Reinwald E. *Trypanosoma congolense* bloodstream forms evade complement lysis in vitro by shedding of immune complexes. Eur J Cell Biol 1990;52:264–9.

[20] Roditi I, Schwarz H, Pearson TW, Beecroft RP, Liu MK, Richardson JP, Buhring HJ, Pleiss J, Bulow R, Williams RO, et al. Procyclin gene expression and loss of the variant surface glycoprotein during differentiation of *Trypanosoma brucei*. J Cell Biol 1989;108:737–46.

JOURNAL OF CLINICAL MICROBIOLOGY, May 2000, p. 1901–1908 Vol. 38, No. 5
0095-1137/00/$04.00+0

Technical Assessment of the Affymetrix Yeast Expression GeneChip YE6100 Platform in a Heterologous Model of Genes That Confer Resistance to Antimalarial Drugs in Yeast

MARTIN E. NAU,[1] LYNDAL R. EMERSON,[2] RODGER K. MARTIN,[3] DENNIS E. KYLE,[3] DYANN F. WIRTH,[2] AND MARYANNE VAHEY[4]*

Henry M. Jackson Foundation for the Advancement of Military Medicine, Rockville, Maryland[1]; Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts[2]; and Division of Experimental Therapeutics[3] and Division of Retrovirology,[4] Walter Reed Army Institute of Research, Washington, D.C.

Received 4 January 2000/Returned for modification 17 February 2000/Accepted 24 February 2000

The advent of high-density gene array technology has revolutionized approaches to drug design, development, and characterization. At the laboratory level, the efficient, consistent, and dependable exploitation of this complex technology requires the stringent standardization of protocols and data analysis platforms. The Affymetrix YE6100 expression GeneChip platform was evaluated for its performance in the analysis of both global (6,000 yeast genes) and targeted (three pleiotropic multidrug resistance genes of the ATP binding cassette transporter family) gene expression in a heterologous yeast model system in the presence and absence of the antimalarial drug chloroquine. Critical to the generation of consistent data from this platform are issues involving the preparation of the specimen, use of appropriate controls, accurate assessment of experiment variance, strict adherence to optimized enzymatic and hybridization protocols, and use of sophisticated bioinformatics tools for data analysis.

A universal challenge to drug therapy is the development of drug resistance. Efforts to understand the molecular mechanisms of the emergence of resistance to drugs span the fields of infectious disease, cancer, and toxicology. The eventuality of drug resistance necessitates the ongoing development of new drugs and interventions. A decade of research has identified a class of genes associated with multidrug resistance (8, 9).

The multidrug resistance genes (*mdr* genes) are part of the ATP binding cassette (ABC) transporter genes in mammalian cells (4, 7, 10). To facilitate the detection of drug resistance and to expedite the development of new drugs, several in vitro model systems have been developed that examine the activity of *mdr* and ABC transporters. One such system is the heterologous yeast model in which the genes *PDR5*, *PDR10*, and *SNQ2*, members of the pleiotropic drug resistance (*pdr*) family in yeast, have been associated with drug resistance (2, 9, 10, 15, 16, 17, 18). Observations that there may be 30 or more genes in yeast that are related by sequence homology to the ABC transporter gene family complicate the association of drug resistance with a particular gene (3). The *Saccharomyces cerevisiae* genome sequencing project revealed 31 ABC genes, which have been classified into six distinct subfamilies based on phylogenetic analysis (3, 7, 14, 19, 20). The *pdr* family is the largest of these subgroups, with 10 members. In total there are 12 ABC genes that have been associated with modulation of resistance to xenobiotics to date. The *PDR5* gene has been linked to resistance to cycloheximide, mycotoxins, and cerulenin, and its product has been found to transport glucocorticoids (2, 3, 4, 10, 13). A second member of the *pdr* group, *SNQ2*, has been found to be linked to resistance to 4-nitroso-quinoline-*N*-oxide, methyl-nitro-nitrosoguanidine, and metal ions such as Na^+, Li^+, and Mn^+ (3, 16, 18). The Δ*snq2* Δ*pdr5* deletion strain exhibits a more pronounced sensitivity to metal ions and other drug substrates (3). *PDR10* is closely related to *PDR5* (65% sequence identity); however, the functional relatedness of these genes remains to be determined. Interestingly, *PDR10* has been found to localize to the cell surface like *PDR5* and *SNQ2* (3, 9).

With the introduction of the Affymetrix yeast expression GeneChip YE6100 platform (YE6100 platform), it has become feasible to plan experiments to simultaneously assess the changes in the expression patterns of not only the pleiotropic drug resistance gene family but also 6,000 yeast genes (5). Previously, Wodicka et al., at Affymetrix, characterized the basic performance characteristics of a prototype for the YE6100 platform to generate a global survey of 6,000 yeast genes (22). This platform was refined and exploited by Cho et al. to survey the complete yeast genome (6). Holstege et al., using an elegant battery of controls, exploited the commercially available YE6100 platform to assess the transcriptional control of yeast cell division (11). Winzeler et al. used a customized gene array platform for direct allelic scanning of the entire yeast genome (21).

To test the practical potential of the commercially available YE6100 platform to address drug resistance, a well-defined heterologous yeast model system was chosen. The expression profiles of two strains of *S. cerevisiae* were evaluated in the presence and absence of the antimalarial drug chloroquine. Strain YPH 499 (499) is wild type and refractory to the drug chloroquine. Strain YHW 1052 (1052) is a mutant with deletions in the *PDR5*, *PDR10*, and *SNQ2* genes and is thus more susceptible to chloroquine. The aim of this paper is to detail the technical aspects of the utilization of the YE6100 platform that are critical to the generation of consistent and reliable gene expression data in the study of drug resistance. The implementation of the methods and protocols presented in this paper will facilitate more intensive efforts to elucidate the

* Corresponding author. Mailing address: Gene Array Laboratory, Walter Reed Army Institute of Research, 1600 East Gude Dr., Rockville, MD 20850. Phone: (301) 251-5058. Fax: (301) 762-7460. E-mail: mvahey@pasteur.hjf.org.

TABLE 1. Cell densities and mRNA yields

Time point[a]	Strain	Cell density (cells/ml) at:			mRNA (μg)	
		Introduction of drug	Harvest			
			Control cultures	Treated cultures	Control cultures	Treated cultures
Early	499	4.3×10^6	1.7×10^7	1.1×10^7	9.9	14.7
	1052	4.5×10^6	1.5×10^7	8.2×10^6	25.5	13.3
Middle	499	4.2×10^6	3.2×10^7	1.4×10^7	6.6	14.4
	1052	4.1×10^6	4.2×10^7	1.1×10^7	8.0	15.4
Late	499	3.0×10^6	3.5×10^7	2.6×10^7	13.1	6.4
	1052	3.2×10^6	3.2×10^7	1.8×10^7	9.4	7.5

[a] Early, 2 h; middle, 3 h; late, 4.5 h. Treated cultures received 1.5, 2.5, and 2.5 mg/ml for the early, middle and late time points, respectively.

details of the molecular interactions involved in the emergence of drug resistance. Two levels of data analysis, the global assessment of functional gene families and the targeted assessment of particular genes, will be addressed to demonstrate the type of information gleaned from each.

MATERIALS AND METHODS

Strains and media. The strains, 1052 and 499, used in this study were the kind gifts of Karl Kuchler of The University and Biocenter of Vienna, Vienna, Austria. The yeast strain 1052 (*Δpdr5*::TR1 *Δsnq2*::*hisG Δpdr10*::*hisG*) was utilized for this study along with its isogenic parental strain 499 (*MATa ade2-101*cc *his3Δ200 leu2-Δ1 lys2-801*am *trp1-Δ1 ura3-52*). Strain 1052 is deficient in three ABC transporters encoded in the *pdr* pathway (*PDR5*, *PDR10*, and *SNQ2*). In strain 1052, the deletion in *PDR5* is from nucleotide (nt) +399 through nt +4456. The deletion in *PDR10* is from nt −90 through nt +4307. The deletion in *SNQ2* is from nt −6 through nt +3899. The 50% inhibitory concentrations of the drug chloroquine are 127 mg/ml for 499 and 50.00 mg/ml for 1052 as determined in nonaerated liquid medium and in solid medium culture. In liquid culture the 50% inhibitory concentrations of the drug chloroquine are 4.75 ± 0.75 mg/ml for 499 and 1.38 ± 0.13 mg/ml for 1052. Starter cultures were taken from colonies lifted from freshly streaked agar plates and grown overnight (to confluence at 2×10^8 cells/ml) at 30°C and 300 rpm in 5 to 10 ml of yeast-peptone-dextrose medium. The 5- to 10-ml starter cultures were diluted into 1,200 ml of prewarmed and aerated yeast-peptone-dextrose medium in a 4-liter flask to a density of 1.5×10^6 cells/ml. Cultures were grown at 30°C and 300 rpm for 2 h or until the cell density reached 3.0×10^6 cells/ml. At this juncture the culture was split into two 600-ml aliquots in two prewarmed 2-liter flasks. Chloroquine was added to the treatment flask to a concentration of 1.5 or 2.5 mg/ml from a 200-mg/ml concentrated stock of chloroquine diphosphate salt (Sigma, St. Louis, Mo.) dissolved in sterile double-distilled water. This solution had a pH of approximately 4.0. An exact volume of sterile double-distilled water, adjusted to the pH of the chloroquine solution, was added to the control flask. Table 1 shows the cell densities from critical points in the growth and treatment of the cultures used in the study. The assay points in the study are defined as early (2 h with or without 1.5 mg of drug per ml), middle (3 h, with or without 2.5 mg of drug per ml), and late (4.5 h, with or without 2.5 mg of drug per ml).

Cell harvesting and preparation of poly(A) RNA. Cultures were harvested identically at three time points: 2, 3, and 4.5 h. It is imperative that all cultures be treated exactly the same during the harvesting procedure. The overnight yeast culture was dispensed into 12 50-ml polypropylene conical tubes (Falcon/Becton Dickinson Labware, Franklin Lakes, N.J.) and centrifuged in a clinical centrifuge for 5 min at 4°C and at 2,000 × *g*. The pellet was resuspended in 5 ml of Tri-Reagent (Molecular Research Center, Woodlands, Tex.), and an equal volume of 400-μm-diameter acid-washed glass beads was added. The mixture was vortexed for 1 min. An additional 20 ml of Tri-Reagent was added to the mixture, and the manufacturer's instructions for the preparation of total RNA were followed. Poly(A) RNA (mRNA) was prepared from total RNA using the Oligotex (Qiagen, Valencia, Calif.) method according to the manufacturer's instructions.

cDNA synthesis. Double-stranded cDNA was synthesized in two steps using the Superscript Choice System (GibcoBRL, Rockville, Md.) and the reverse transcription primer T7-$(dt)_{24}$ [5′GGCCAGTGAATTGTAATACGACTCACT ATAGGGAGGCGG$(T)_{24}$ 3′] (GENSET Corp., LaJolla, Calif.). First-strand synthesis was carried out in a 20-μl reaction mixture. Approximately 3.0 μg of mRNA was annealed to 7 μg of T7-$(dt)_{24}$ primer at 70°C for 10 min. Reverse transcription was carried out at 37°C for 1 h in a mixture with final concentrations of 50 mM Tris-HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$, 10 mM dithiothreitol, 500 μM each dATP, dCTP, dGTP, and dTTP, and 20,000 to 30,000 U of Superscript II reverse transcriptase per ml, and the reaction was terminated by placing the tube on ice. Second-strand synthesis was carried out in 150 μl, incorporating the entire 20-μl first-strand reaction mixture and a 130-μl second-strand reaction mixture for final concentrations of 25 mM Tris-HCl (pH 7.5), 100 mM KCl, 5 mM $MgCl_2$, 10 mM $(NH_4)_2SO_4$, 0.15 mM β-NAD^+, 250 μM each dATP, dCTP, dGTP, and dTTP, 1.2 mM dithiothreitol, 65 U of DNA ligase per ml, 250 U of DNA polymerase I per ml, and 13 U of RNase H per ml. The mixture was incubated at 16°C for 2 h, whereupon 2 μl of T4 DNA polymerase at 5 U/μl was added and the incubation was continued at 16°C for 5 min. To terminate the reaction, 10 μl of 0.5 M EDTA was added. The cDNA was purified using phenol-chloroform-isoamyl alcohol (24:23:1) saturated with 10 mM Tris-HCl (pH 8.0)–1 mM EDTA (AMBION, Inc., Austin, Tex.). The purified cDNA was precipitated with 5 M ammonium acetate and absolute ethanol at −20°C for 20 min. The pellet was resuspended in 7 to 9 μl of RNase-free water to achieve a final concentration of between 0.25 and 0.65 μg/μl.

In vitro transcription and fluorescent labeling. Synthesis of biotin-labeled cRNA was carried out by in vitro transcription using the MEGAscript T7 *In Vitro* Transcription Kit (AMBION, Inc.). According to the manufacturer's instructions, 0.4 to 1.0 μg of double-stranded cDNA was placed in a 20-μl reaction mix, at room temperature, containing Ambion 1× reaction buffer and enzyme mix (proprietary). The labeling mix consisted of 7.5 mM ATP, 7.5 mM GTP, 5.6 mM UTP, 1.9 mM biotinylated UTP (ENZO Diagnostics, Farmingdale, N.Y.), 5.6 mM CTP, and 1.9 mM biotinylated CTP (ENZO). The reaction mixture was incubated at 37°C for 5 h. The biotin-labeled cRNA was purified using RNeasy spin columns (Qiagen) according to the manufacturer's protocol. The biotin-labeled cRNA was fragmented in a 40-μl reaction mixture containing 40 mM Tris-acetate (pH 8.1), 100 mM potassium acetate, and 30 mM magnesium acetate, incubated at 94°C for 35 min, and then put on ice. One microliter of the intact biotin-labeled cRNA and 2 μl of the fragmented sample were run on a 1% agarose gel to evaluate both the yield and size distribution of the intact and fragmented products.

Hybridization, staining, and scanning of the GeneChip. The biotin-labeled and fragmented cRNA was hybridized to the YE6100 Yeast GeneChip array (Affymetrix, Santa Clara, Calif.) according to the manufacturer's instructions. Briefly, a 220-μl hybridization solution of 1 M NaCl, 10 mM Tris (pH 7.6), 0.005% Triton X-100, 50 pM control oligonucleotide B2 (5′ bioGTCAAGATG CTACCGTTCAG 3′) (Affymetrix), control cRNA (Bio B [150 pM], Bio C [500 pM], Bio D [2.5 nM], and Cre X [10 nM]) (American Type Tissue Collection, Manassas, Va., and Lofstrand Labs, Gaithersburg, Md.), 0.1 mg of herring sperm DNA per ml, and 0.05 μg of the fragmented labeled sample cRNA per μl was heated to 95°C, cooled to 40°C, and clarified by centrifugation before being applied to each of the four subarrays (A, B, C, and D) that comprise the YE6100 Yeast GeneChip platform. Hybridization was at 40°C in a rotisserie hybridization oven (model 320; Affymetrix) at 60 rpm for 16 h. Following hybridization, the GeneChip arrays were washed 10 times at 25°C with 6× SSPE-T buffer (1 M NaCl, 0.006 M EDTA, 0.06 M Na_3PO_4, 0.005% Triton X-100, pH 7.6) using the automated fluidics station protocol. GeneChip arrays were incubated at 50°C in 0.5× SSPE-T for 20 min at 60 rpm in the rotisserie oven and then stained for 15 min room temperature and 60 rpm with streptavidin phycoerythrin (Molecular Probes, Inc., Eugene, Oreg.) stain solution at a final concentration of 10 μg/ml in 6× SSPE-T buffer and 1.0 mg of acetylated bovine serum albumin (Sigma) per ml. The GeneChip arrays were washed twice at room temperature with 6× SSPE-T buffer and then scanned with a GeneArray Scanner (Hewlett-Packard, Santa Clara, Calif.), controlled by GeneChip 3.1 software (Affymetrix).

Assay monitoring and controls. The TEST 1 GeneChip (Affymetrix) was used according to the manufacturer's instructions to assess critical features of the mRNA preparations and the cDNA generated from the yeast strains and to evaluate the stringency of staining and hybridization. In addition, a battery of three types of GeneChip controls present on the TEST 1 GeneChip and on each of the four arrays in the YE6100 GeneChip set were employed according to the manufacturer's instructions. Details of the use and performance of these critical controls are given in Results. A method of mathematical scaling was employed by the GeneChip 3.1 software (Affymetrix) to normalize the fluorescence signal from each probe cell on each GeneChip and thus facilitate the reliable comparison of data from independent experiments.

Data analysis algorithm for the assessment of variance. The Affymetrix raw data set was scrutinized to eliminate any transcripts with fewer than 50% of probe cells contributing to the data. Subsequently, the first step in raw data mining for the assessment of variance captured all gene transcripts that were present on both GeneChips being compared (PP data set). The second step required that a decision be made to define what degree of change would be considered significant. We chose to approach this issue objectively, using a distribution analysis of the complete PP data set which defined a mean for the population of values and subsequently determined quartile percentages of 25, 50, and 75% above and below that mean. For the assessment of variance, outliers were defined as values exceeding the mean by 10-fold and were eliminated from the data set. When the PP data set was examined in this way, a value of 3.0-fold was determined to be the cutoff for a reliable change in expression. The value of 3.0-fold was applied to all subsequent analyses. Variances between GeneChips (intraexperimental variance) and between independent mRNA targets (interex-

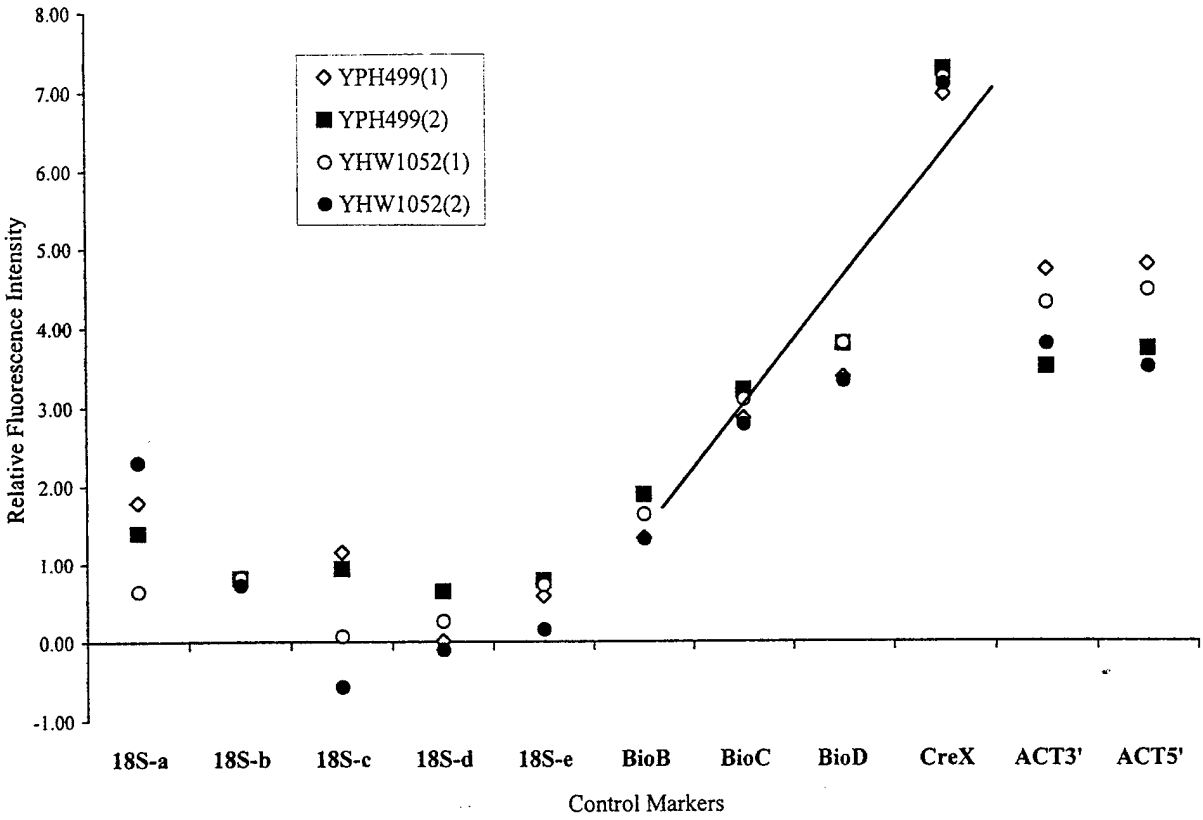


FIG. 1. Performance of the battery of GeneChip controls with two independent preparations of templates from strains 1052 and 499. The ordinate shows the relative fluorescence values for each of the control markers listed on the abcissa. The linear regression r^2 value for the standard curve generated by the Bio B, Bio C, Bio D, and Cre X markers is 0.86.

perimental variance) were assessed by scoring the percentage of PP transcripts that exhibit no change relative to the total number of PP transcripts.

Data analysis algorithm for interrogation. For experiments in which differences in expression profiles between the drug-treated and untreated yeast strains were examined, the data analysis captured data from genes that were present in both cases (PP data set), as well as genes present in one case and absent in the other (PA or AP data set). All values above the 3.0-fold cutoff were included in the analysis of experimental expression profiles. The experimental design employed the analysis of data from the untreated control as a baseline for comparison to the treated strain in all cases. The cumulative fold change for the expression of all genes in a particular functional family was the sum of the levels of change of gene expression, using the values for the untreated strain as the control.

Bioinformatics analyses. GeneSpring version 2.1 (Silicon Genetics, San Carlos, Calif.) was used to derive global trends in the expression profiles and to specifically assess the expression patterns of the *pdr* gene targets. We used the temporal analysis of all of the raw data from the Affymetrix platform normalized to a single mean by the GeneSpring software.

RESULTS

Consistent cell harvests and mRNA yields. Table 1 shows the yields of cells and of mRNA across the three time points of the experiment and at the two concentrations of chloroquine used in the study. The amounts of cells harvested were comparable and equivalent at all time points.

Assessment of GeneChip performance. A battery of controls was used for all experiments. Three types of GeneChip controls are present on the TEST 1 GeneChip and on each of the four GeneChips in the YE6100 set. The first set of controls consists of four synthetically generated plasmid templates that are subjected to reverse transcription to incorporate fluorescent label according to the manufacturer's instructions (Affymetrix). These four cRNA templates, Bio B, Bio C, Bio D, and Cre X, are mixed in a cocktail to generate final concentrations of 150 pM, 500 pM, 2.5 nM, and 10 nM, respectively. These concentrations generate a standard curve and can thus be used to standardize interexperimental variation and efficiency of cDNA synthesis and labeling and to provide the dynamic range of the assay. Ultimately, the standard curve generated by these templates can be used to quantitate the level of RNA expression for a given gene on a per-cell basis. The second set of controls used on the GeneChip assesses the efficiency of cDNA synthesis by quantitating the amounts of 3′ and 5′ portions of target sequences generated during cDNA synthesis by assessing the expression of the yeast actin gene. Optimal synthesis reactions will generate equivalent amounts of signal in the 3′ and 5′ prime targets. The third set of GeneChip controls involves the evaluation of the integrity of the mRNA preparation used in the analysis and reports the GeneChip-based determination of equivalent amounts of mRNA used in the test. This is achieved by the assessment of the 18S rRNA gene expression profile, which is divided on the GeneChip into five sets of probe cells or segments (a through e).

The results of the analysis of TEST 1 GeneChip controls for two independent evaluations of strains 499 and 1052 are shown in Fig. 1. The ordinate indicates the relative fluorescence intensity reported by the GeneArray Scanner. The data from the 18S rRNA series show less than a twofold range in segment a and no significant difference in segment b, c, or d, except for the 1052 data point, which is less than onefold lower in segment c. This data set supports the hypothesis that equivalent amounts of mRNA were used in the cDNA reaction in preparation for GeneChip analysis. Also shown in Fig. 1 are the

TABLE 2. Descriptive statistical analysis of the GeneChip control platform

Control type	n^a	Relative fluorescence intensity[b] with strain:	
		1052	499
Bio B, Bio C, Bio D, and Cre X	20	4.13 ± 0.53 (r^2 = 0.83)	3.27 ± 0.56 (r^2 = 0.81)
18S ribosomal genes (a–e)	40	0.98 ± 0.38	1.04 ± 0.43
Yeast actin, 3′ and 5′	36	2.74 ± 0.55	3.04 ± 0.55

[a] Total number of data points in the analysis.
[b] Data are expressed as the median ± the standard error. Regression values for the standard curve are for linear regression calculations.

results of the assessment of 3′ and 5′ segments of the actin gene expression. There is no significant difference between the fluorescence values for the 3′ end of the yeast actin gene and for the 5′ end of the yeast actin gene in this experiment. This data set indicates an optimal yield from the cDNA synthesis reaction. The manufacturer (Affymetrix) suggests that the yield of 3′ product may vary by as much as fourfold. In our hands, optimization of the cDNA synthesis step routinely yielded less than a 0.5-fold difference between 3′ and 5′ segments.

The standard curve generated by the synthetic templates Bio B, Bio C, Bio D, and Cre X is shown in Fig. 1. The curve has an r^2 value of 0.86 and was remarkably consistent between strains, between GeneChips, and for two independent template preparations. Table 2 summarizes data on the performance of the battery of the three sets of controls that were generated by between 20 and 40 independent GeneChip assessments. A descriptive statistical analysis of the data set shows stringent inter- and intraexperimental consistency.

Assessment of assay variance. Table 3 presents data on the results of two independent expression profiles for each strain, 499 and 1052, in the absence of drug. These data were generated using one of the four GeneChips that comprise the complete YE6100 GeneChip platform (GeneChips A through D). In each case an independent growth and harvest of yeast cells followed by an independent preparation of GeneChip-ready template was carried out. Genes were scored as being present in both sets of data (PP), exhibiting no change in expression between the two sets of data, having increased or decreased, and, finally, having increased or decreased by threefold. For strain 1052, the total number of PP genes was 1,450, of which 32 increased by threefold, 116 decreased by threefold, and 1,302 (89%) remained unchanged, thus generating a variance between the two runs of 10.2%. For strain 499, the total number of PP genes was 1,439, of which 72 increased by threefold, 153 decreased by threefold, and 1,214 (84%) remained unchanged, thus generating a variance between the two runs of 15.6%. To further reduce these levels of interexperimental variance, the original culture was split into two cultures and reassessed for percentage of variance. As a result of splitting the original culture in this way, rather than growing two side-by-side cultures, the variance was reduced to zero for both strains, since there were no genes that changed greater than threefold between the two runs.

Global expression profiles of strains 1052 and 499 in the presence and absence of chloroquine. Shown in Fig. 2 and 3 are the results of a global survey of the 6,000 genes on the YE6100 GeneChip platform as assessed in strains 1052 and 499, respectively, in the presence and absence of the drug chloroquine and at each of the three time points and two drug concentrations used in the study. The control in each case was the value from the strain in the absence of the drug. Cumulative fold change values for the functional families are arrived at by simple summation of the levels of change from the control for each gene in a functional family.

As compared with the middle and late time points, the early time points for both 1052 and 499 exhibit a lower level of expression, with some increase in genes associated with membranes in strain 499. At the middle time point, however, both strains exhibit an increase in gene expression, with few genes showing a decrease. Genes associated with membranes, metabolism, and ribosomes showed the most increase in strain 1052 at the middle time point. The levels of the cumulative increase in expression were 2- to 10-fold higher in strain 499 at the middle time point. Increases in the expression of genes associated with membranes, metabolism, and ribosomes were similar in pattern but greater in magnitude to the changes at this time point in strain 1052. The most dramatic change occurred in strain 499 at the middle time point in the increase in expression of genes associated with synthetic pathways. In strain 1052, the late time point data set was dominated by a large decrease in the expression of genes associated with membranes. In contrast to the case for the two earlier time points, most expression levels were reduced in strain 1052 at the late time point. The expression of genes in strain 499 was also decreased at the late time point compared with the two earlier time points. The largest decline in expression was in the genes associated with translation and transcription.

Targeted expression profiles of the *pdr* genes *PDR5*, *PDR10*, and *SNQ2* in strains 1052 and 499 in the presence and absence of chloroquine. Figure 4 shows the expression profiles at three time points and in the absence or the presence of two different concentrations of the antimalarial drug chloroquine. The expression of the gene *PDR5* was decreased in the 1052 mutant strain in the presence and absence of the drug. In contrast, the expression of the gene *PDR10* was increased in strain 1052 in the presence and the absence of chloroquine. The expression

TABLE 3. Calculation of variance for two independently grown and tested samples of each strain

Strain	No. of genes present in both profiles	No. (%) of genes:						Variance	
		With no change	That increase	That decrease	With <3-fold or no change	That increase ≥3-fold	That decrease ≥3-fold	No.	%
1052	1,450	786 (54.2)	394 (27.2)	270 (18.6)	1,302 (89.8)	32 (2.2)	116 (8.0)	148	10.2
499	1,439	690 (47.9)	421 (29.3)	328 (22.8)	1,214 (84.4)	72 (5.0)	153 (10.6)	225	15.6

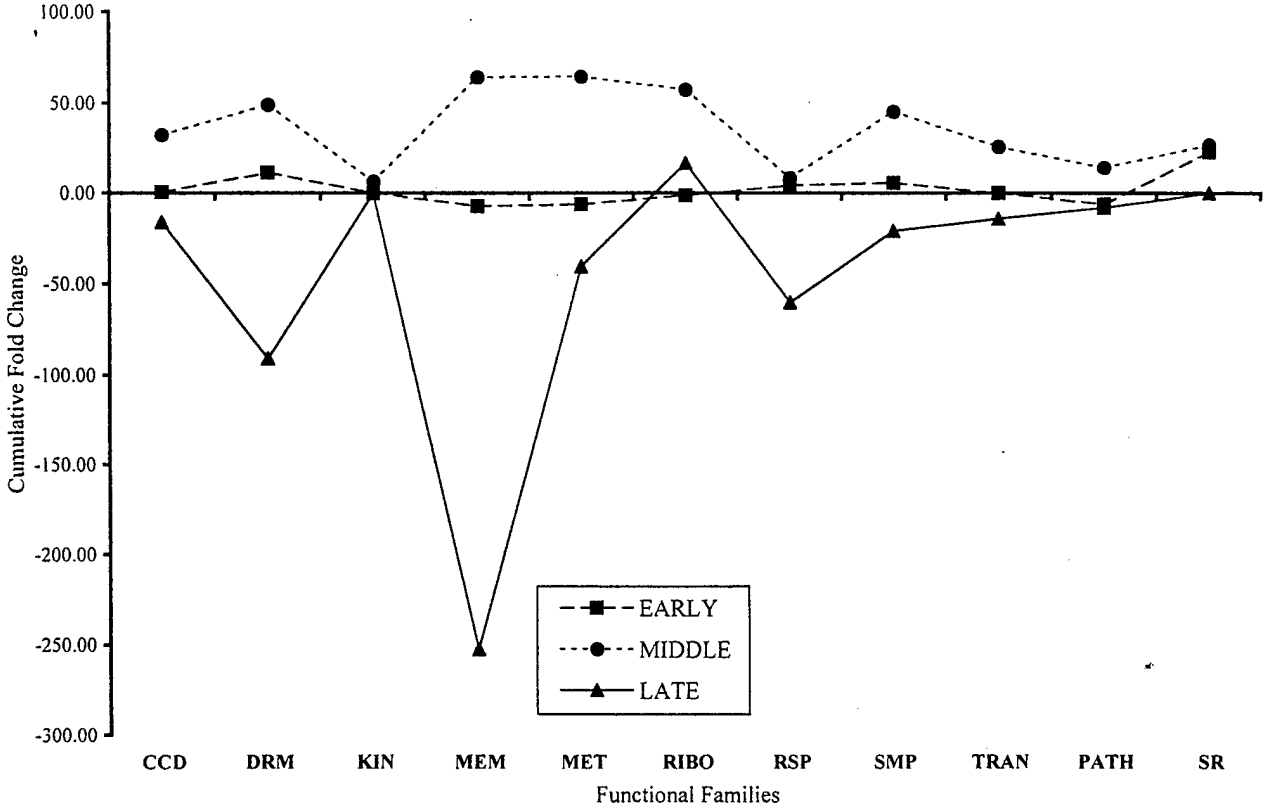


FIG. 2. Cumulative change of gene expression levels in the mutant strain 1052 in the presence of chloroquine. The ordinate shows the cumulative fold changes for the expression levels of genes categorized by the functional family designation shown on the abcissa. The functional families are cell cycle and division proteins (CCD), drug resistance membrane proteins (DRM), kinases (KIN), membrane proteins (MEM), metabolic pathway proteins (MET), ribosomal proteins (RIBO), respiratory chain proteins (RSP), synthetic metabolic pathways (SMP), transcription and translation proteins (TRAN), pathology-related proteins (PATH), and stress-related proteins (SR). The expression level of genes in the untreated sample is used to determine the baseline for the degree of change of gene expression. The profiles for the early time point, the middle time point, and the late time point are shown.

of the gene *SNQ2* was moderate but level in strain 1052 in the presence of drug and moderate with a minor increase in slope in the absence of the drug. The wild-type strain 499 exhibited an increase in the expression levels of *PDR5* in the presence of drug but not in the absence of drug. In the absence of drug, the expression of the gene *PDR5* was moderate and level across all time points. The expression levels of *PDR10* and *SNQ2* in strain 499 remained low and level in both the presence and absence of the drug.

DISCUSSION

Template preparation. Several approaches to the extraction of total RNA and the subsequent preparation of mRNA are currently available. We found that the combination of two commercially available kits, the Tri-Reagent and Qiagen Oligotex kits, gave the most dependable results with yeast. The most critical aspects of the preparation of template for the Affymetrix GeneChip YE6100 platform are the quality of the mRNA and the degree to which it is representative of the biological nature of the sample. To ensure a representative sample, it is imperative to standardize the growth and handling of the yeast cultures. Holstege et al. first suggested that the attention to detail involved in the growth and harvest of yeast cultures for expression profiling was critical to the dependability of the data generated (11). We confirm and extend that observation by emphasizing the added importance of standardizing the treatment of these strains with the drug chloroquine and minimizing experimental variance by splitting single cultures for drug treatment. It is imperative to ascertain the phenotypes of the wild-type and mutant strains in the presence of a drug prior to the characterization of the expression profiles generated as a result of treatment with that drug.

Quality control and assessment. The Affymetrix GeneChip YE6100 is exquisitely sensitive and necessitates the use of powerful controls to assure that all aspects of the procedure are consistent and reliable. Of the four types of controls available for expression profiling, using the Affymetrix GeneChip YE6100, we chose to apply three. The only control that we did not utilize involved the addition of synthetic total RNA template to the RNA samples extracted from the yeast strains. Instead, we chose to use data from the 3′ and 5′ ends of the yeast actin gene as a more accurate and less intrusive measure of the yield, quality, and representative nature of the mRNA. The data generated by these controls result directly from the sample tested and are not enhanced or quenched by the presence of artificial template.

We have determined that the battery of three controls that we routinely employ are essential to the interpretation, consistency, and reliability of expression profiling experiments. Perhaps the most powerful of the sets of controls is the standard curve generated by the synthetic templates Bio B, Bio C, Bio D, and Cre X. These data points offer the investigator the power to express GeneChip data on a semiquantitative level. The 18S ribosomal protein series and the yeast actin 3′ and 5′

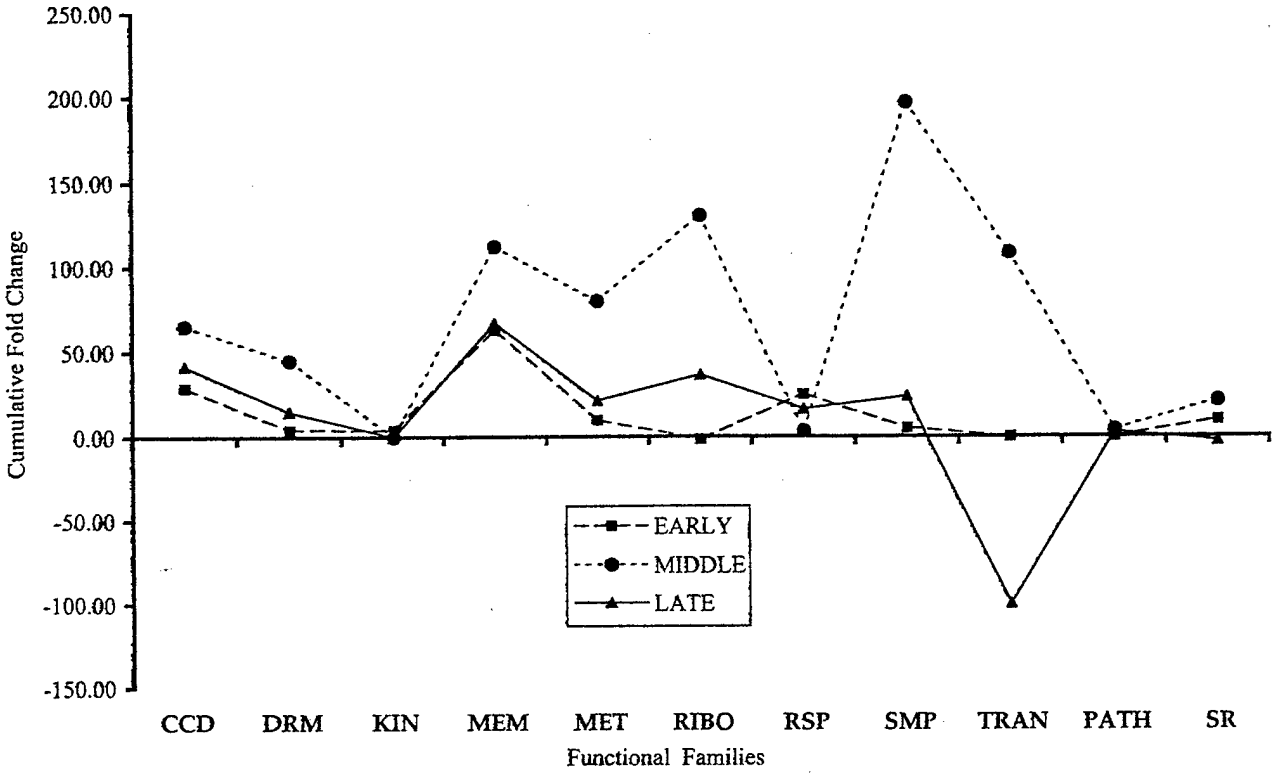


FIG. 3. Cumulative change of gene expression levels in the wild-type strain 499 in the presence of chloroquine. The ordinate shows the cumulative fold change for the expression levels of genes categorized by the functional family designation shown on the abcissa. The functional families are cell cycle and division proteins (CCD), drug resistance membrane proteins (DRM), kinases (KIN), membrane proteins (MEM), metabolic pathway proteins (MET), ribosomal proteins (RIBO), respiratory chain proteins (RSP), synthetic metabolic pathways (SMP), transcription and translation proteins (TRAN), pathology-related proteins (PATH), and stress-related proteins (SR). The expression level of genes in the untreated sample is used to determine the baseline for the degree of change of gene expression. The profiles for the early time point, the middle time point, and the late time point are shown.

end targets provide critical information on the preparation of the RNA and on the representative quality of the cDNA subsequently produced. The fact that all of these controls reside on each GeneChip further supports and ensures the generation of dependable data both within and between experiments. Most importantly, remarkably low levels of intraexperimental variance can be achieved, despite the enormous number of complex steps involved in generating an expression profile, by faithful attention to optimized laboratory protocols and by the vigilant use of the battery of GeneChip controls.

Interpretation of GeneChip expression profiles. We employed a well-characterized heterologous yeast model to assess the impact of the drug chloroquine on the yeast *pdr* genes *PDR5*, *PDR10*, and *SNQ2*. We assessed the expression profile data on two levels: (i) the global analysis of cumulative changes in expression of genes classified into broad functional families and (ii) the targeted expression analysis of the three *pdr* genes across the three time points and two drug concentrations used in the study. Jelinsky and colleagues used the global assessment of expression profiles to assess changes in gene expression in yeast in response to alkylating agents (12). Alon and colleagues employed targeted expression and cluster analysis to define expression patterns in colon tumors (1).

The assessment of the global alterations in expression profiles of broadly defined functional families in each of the strains in the presence of drug clearly identifies that in the mutant, the functional family most significantly affected by the drug is the membrane protein group. Strain 1052 exhibits a 250-fold reduction in the cumulative gene expression in the membrane protein group. The functional family of drug resistance-related membrane proteins is also reduced in cumulative gene expression by 75-fold. In contrast to these observations, the wild-type strain exhibits an increase in the expression of membrane-associated proteins and, most significantly, in proteins involved with the processes of transcription and translation. By the late time point, the wild-type strain exhibits a 100-fold decrease in the expression of proteins related to transcription and translation. Clearly these two strains respond with distinct strategies to the presence of drug. The assessment of the degree of cumulative change in the expression profiles of broadly defined functional families of genes can be readily made from the data reported by the Affymetrix GeneChip YE6100 platform. This information is most useful in suggesting the focus of further data mining to elucidate the specifics of a biological pathway affected by the drug.

The GeneSpring bioinformatics platform commercially available from Silicon Genetics interrogates the Affymetrix GeneChip YE6100 data in a significantly more powerful way. This tool allows for the identification of the patterns and magnitude of expression of any single gene assessed by the Affymetrix GeneChip YE6100 over the course of the study. The expression profile of individual targeted genes as well as the patterns or clusters of related genes can also be elucidated by the analysis. In the model system employed in this study, the promoter region of the *PDR10* gene was disrupted. An unchanged or reduced expression of this gene might be predicted as a result of this deletion. The expression profiles derived by GeneSpring analysis of the *PDR10* gene in the mutant strain

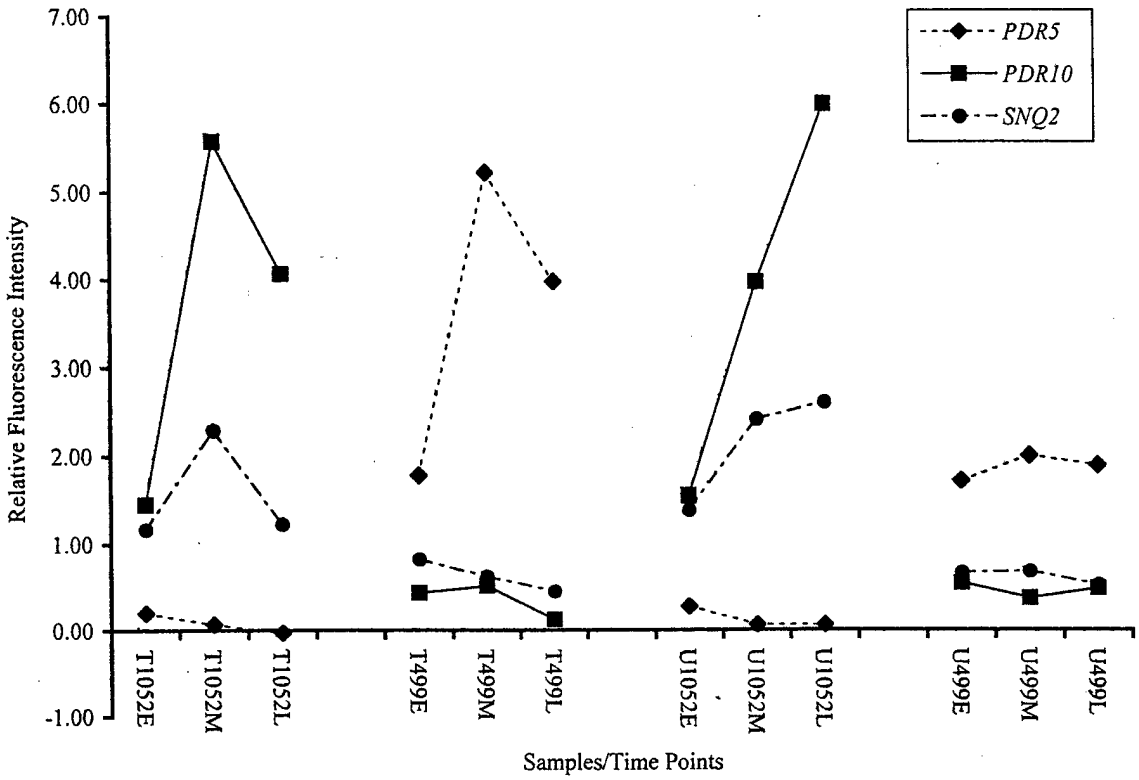


FIG. 4. Expression profiles of the yeast *pdr* genes *PRD5*, *PDR10*, and *SNQ2*. The ordinate shows the relative fluorescence intensity for (i) each of the study time points (early [E], middle [M], and late [L]), (ii) the two experimental treatments (drug treated [T] and untreated [U], and (iii) each of the two strains (1052 and 499), as shown on the abcissa.

1052 exhibit an unexpectedly high level of expression in both the presence and absence of chloroquine. Several explanations for this observation can be proposed.

The elevated levels of the mutant *PDR10* gene expression may reflect the bias of the GeneChip to assess the 3′ region of a gene. It is important to take into account that the Affymetrix GeneChip YE6100 platform interrogates 25-mer regions that cover the last 600 bp of the 3′ end of the gene (5). This region is distal to the deletion made at the 5′ promoter region of the *PDR10* gene. Alternatively, there may be a difference in the efficiency of the promoter region, or in the stability or rate of turnover of the gene product, in the mutant as compared to that of the intact gene in the wild-type strain. In the wild-type strain, there is an increase in the production of *PDR5* in response to drug treatment, while the *PDR10* and *SNQ2* expression levels remain moderate and unchanged, respectively. This pattern may reflect the specificity of the *PDR5* response to the drug chloroquine in this strain (9). In contrast, expression levels of *PDR5* and *SNQ2* in the mutant strain show little or no response to the presence of the drug. Mechanistic explanations of the biological function of the gene products of *PDR5*, *PDR10*, and *SNQ2* in the mutant and wild-type strains warrant further investigation. These observations show the complexity of the interpretation of expression profile data and underscore the necessity of ascertaining, by an independent assessment, information on the functional status of a gene target.

In summary, the utilization of optimized laboratory protocols, monitored by stringent controls, generates a powerful data set from the Affymetrix Expression GeneChip platform. The interpretation of the patterns and magnitudes of expression profiles represented in the data set requires the application of bioinformatics tools and a fundamental knowledge of the model being examined. The power of the method resides in the sensitivity, accuracy, and speed with which the expression of over 6,000 genes in response to experimental conditions can be simultaneously assessed. Confirmation of the trends observed in the data generated by expression profiling serves as a point of departure for further analysis of gene function and thus of the molecular mechanisms of drug action.

ACKNOWLEDGMENTS

We thank Karl Kuchler for the yeast strains YHW 1052 and YPH 499 used in this study, Anthony Lailin of Silicon Genetics and Brian Shimada and Mark Hurt of Affymetrix for their technical advice, and Deborah L. Birx and Nelson L. Michael for helpful discussions.

This work was supported in part by Cooperative Agreement no. DAMD17-93-V-3004 between the U.S. Army Medical Research and Materiel Command and the Henry M. Jackson Foundation for the Advancement of Military Medicine.

REFERENCES

1. **Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and S. J. Levine.** 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96:6745–6750.
2. **Balzi, E., M. Wang, S. Leterme, L. Van Dyck, and A. Goffeau.** 1994. *PDR5*, a novel yeast multidrug resistance conferring transporter controlled by the transcription regulator PDR1. J. Biol. Chem. 269:2206–2214.
3. **Bauer, B. E., H. Wolfger, and K. Kuchler.** 1999. Inventory and function of yeast ABC proteins: sex, stress, pleiotropic drug and heavy metal resistance. Biochim. Biophys. Acta 1461:217–236.
4. **Bissinger, P. H., and K. Kuchler.** 1994. Molecular cloning and expression of the *Saccharomyces cerevisiae* STS1 gene product. A yeast ABC transporter conferring mycotoxin resistance. J. Biol. Chem. 269:4180–4186.

5. **Chee, M., R. Yang, E. Hubbell, A. Berno, X. C. Huang, D. Stern, J. Winkler, D. J. Lockhart, M. S. Morris, and S. P. A. Fodor.** 1996. Accessing genetic information with high density DNA arrays. Science **274**:610–614.
6. **Cho, R. J., M. Fromont-Racine, L. Wodicka, B. Feierbach, T. Streams, P. Legrain, D. J. Lockhart, and R. W. Davis.** 1998. Parallel analysis of genetic selections using whole genome oligonucleotide arrays. Proc. Natl. Acad. Sci. USA **95**:3752–3757.
7. **Decottignies, A., and A. Goffeau.** 1997. Complete inventory of the yeast ABC proteins. Nat. Genet. **15**:137–145.
8. **Delling, U., M. Raymond, and E. Schurr.** 1998. Identification of *Saccharomyces cerevisiae* genes conferring resistance to quinoline ring containing antimalarial drugs. Antimicrob. Agents Chemother. **42**:1034–1041.
9. **Egner, R., F. E. Rosenthal, A. Kralli, D. Sanglard, and K. Kuchler.** 1998. Genetic separation of FK506 susceptibility and drug transport in the yeast *PDR5* ATP binding cassette multidrug resistance transporter. Mol. Biol. Cell **9**:523–543.
10. **Hirata, D., K. Yano, K. Miyahara, and T. Miyakawa.** 1994. *Saccharomyces cerevisiae YDR1*, which encodes a member of the ATP-binding cassette (ABC) superfamily, is required for multidrug resistance. Curr. Genet. **26**: 285–294.
11. **Holstege, F. C. P., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. S. Young.** 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95**:717–728.
12. **Jelinsky, S. A., and L. Samson.** 1999. Global response of *Saccharomyces cerevisiae* to an alkylating agent. Proc. Natl. Acad. Sci. USA **95**:1486–1491.
13. **Kralli, A., S. P. Bohen, and K. R. Yamamoto.** 1995. LEM1, an ATP-binding-cassette transporter, selectively modulates the biological potency of steroid hormones. Proc. Natl. Acad. Sci. USA **92**:4701–4705.
14. **Kuchler, K., and R. Egner.** 1997. Unusual protein secretion and translocation pathways in yeast: implication of ABC transporters, p. 49–85. *In* K. Kuchler, A. Rubartelli, and B. Holland (ed.), Unusual secretory pathways: from bacteria to man. Landes Bioscience, Austin, Tex.
15. **Mahe, Y., Y. Lemoine, and K. Kuchler.** 1996. The ATP binding cassette transporters *Pdr5* and *Snq2* of *Saccharomyces cerevisiae* can mediate transport of steroids in vivo. J. Biol. Chem. **271**:25167–25172.
16. **Miyahara, K., M. Mizunuma, D. Hirata, E. Tsuchiya, and T. Miyakawa.** 1996. The involvement of the *Saccharomyces cerevisiae* multidrug resistance transporters Pdr5p and Snq2p in cation resistance. FEBS Lett. **399**:317–320.
17. **Ruetz, S., U. Delling, M. Brault, E. Schurr, and P. Gros.** 1996. The pmdr1 gene of *Plasmodium falciparum* confers cellular resistance to antimalarial drugs in yeast cells. Proc. Natl. Acad. Sci. USA **93**:9942–9947.
18. **Servos, J., E. Haase, and M. Brendel.** 1993. Gene *SNQ2* of *Saccharomyces cerevisiae*, which confers resistance to 4-nitroquinoline-N-oxide and other chemicals, encodes a 169 kDa protein homologous to ATP-dependent permeases. Mol. Gen. Genet. **236**:214–218.
19. **Taglicht, D., and S. Michaelis.** 1998. *Saccharomyces cerevisiae* ABC proteins and their relevance to human health and disease. Methods Enzymol. **292**: 130–162.
20. **Wendler, F., H. Bergler, K. Prutej, H. Jungwith, G. Zisser, K. Kuchler, and G. Hogenauer.** 1997. Diazaborine resistance in the yeast *Saccharomyces cerevisiae* reveals a link between YAP1 and the pleiotropic drug resistance genes *PDR1* and *PDR3*. J. Biol. Chem. **272**:27091–27098.
21. **Winzeler, E. A., D. R. Richards, A. R. Conway, A. L. Goldstein, S. Kalman, M. J. McCullough, J. H. McCusker, D. A. Stevens, L. Wodicka, D. J. Lockhart, and R. W. Davis.** 1998. Direct allelic variation scanning of the yeast genome. Science **281**:1194–1197.
22. **Wodicka, L. H. Dong, M. Mittman, M. H. Ho, and D. J. Lockhart.** 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat. Biotechnol. **15**:1359–1367.

ELSEVIER

Molecular and Biochemical Parasitology 000 (2000) 000–000

MOLECULAR AND BIOCHEMICAL PARASITOLOGY

www.parasitology-online.com.

Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A–T rich genomes

Anusha Munasinghe [a], Swati Patankar [a], Brian P. Cook [b], Steve L. Madden [b], Rodger K. Martin [c,1], Dennis E. Kyle [c], Azadeh Shoaibi [d], Leda M. Cummings [d], Dyann F. Wirth [a,*]

[a] *Department of Immunology and Infectious Diseases, Harvard School of Public Health, Harvard University, Building 1, Room 704, 665 Huntington Ave, Boston MA 02115, USA*

[b] *Genzyme Molecular Oncology, Genzyme Corporation, Framingham, MA 01701-9322, USA*

[c] *Department of Parasitology, Division of Experimental Therapeutics, Walter Reed Army Institute of Research, Silver Springs, MD 20910, USA*

[d] *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA*

Received 21 June 2000; received in revised form 26 September 2000; accepted 27 November 2000

Abstract

The advent of high-throughput methods for the analysis of global gene expression, together with the Malaria Genome Project open up new opportunities for furthering our understanding of the fundamental biology and virulence of the malaria parasite. Serial analysis of gene expression (SAGE) is particularly well suited for malarial systems, as the genomes of Plasmodium species remain to be fully annotated. By simultaneously and quantitatively analyzing mRNA transcript profiles from a given cell population, SAGE allows for the discovery of new genes. In this study, one reports the successful application of SAGE in *Plasmodium falciparum*, 3D7 strain parasites, from which a preliminary library of 6880 tags corresponding to 4146 different genes was generated. It was demonstrated that *P. falciparum* is amenable to this technique, despite the remarkably high A–T content of its genome. SAGE tags as short as 10 nucleotides were sufficient to uniquely identify parasite transcripts from both nuclear and mitochondrial genomes. Moreover, the skewed A–T content of parasite sequence did not preclude the use of enzymes that are crucial for generating representative SAGE libraries. Finally, a few modifications to DNA extraction and cloning steps of the SAGE protocol proved useful for circumventing specific problems presented by A–T rich genomes.

Keywords: *Plasmodium falciparum*; Malaria; Serial analysis of gene expression; Genomics

1. Introduction

The malarial parasite, *Plasmodium falciparum*, infects approximately 250 million people worldwide and kills almost 2 million of these individuals, mostly young children in Africa, annually [1]. The pathogen's success can be largely attributed to its ability to effectively evade host immunity, develop rapid resistance to anti-malarial compounds, and complete a complex life cycle in both the human host and mosquito vector. With the absence of successful vaccination and a paucity of chemotherapeutic drugs, it is evident that insight into parasite biology is vital for developing knowledge-based strategies against a disease that plagues most of the developing world.

With this aim in mind, global malaria initiatives have launched the Malaria Genome Project [2–4] in order to sequence the entire genome of the 3D7 strain of *P. falciparum*. By defining every single gene in the proto-

Abbreviations: AE, anchoring enzyme; BLAST, basic local alignment search tool; bp, base pairs; DMSO, dimethylsulfoxide; LoTE, low Tris-EDTA solution; ORF, open reading frame; PBS, phosphate buffered saline; PCI, phenol:chlorofom:isoamyl alcohol in a 25:24:1 ratio; PCR, polymerase chain reaction; SAGE, serial analysis of gene expression; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis; µl, microliter.

* Corresponding author. Tel.: +1-617-4321621; fax: +1-617-4324766.

E-mail address: dfwirth@hsph.harvard.edu (D.F. Wirth).

[1] Present address: U.S. Army Medical Research and Material Command, Fort Detrick, Federick, MD 21702-5012, USA.

0166-6851/00/$ - see front matter
PII: S0166-6851(00)00378-9

zoan parasite, the Genome Project seeks to uncover virulence factors as well as new targets for vaccine and drug development. The genome spans approximately 24.6 Mb, consists of 14 chromosomes, and is highly A–T rich (70–80% A–T content). At least 80% of the estimated 7000 genes are at least partially sequenced. Chromosomes 3 [5] and 2 [6] were completely sequenced recently; preliminary analysis of these predicts that 60% of coding regions in the malaria genome will have unknown function. The percentage of unidentified open reading frames (ORFs) in the parasite is significant (compared to other genomes) and may reflect *Plasmodium*'s unique requirement for novel genes during host–parasite interactions, or evasion of immune and drug pressure.

Relating genomic sequence to function and ultimately malarial biology is the next logical step. One approach involves investigating transcriptional profiles in *Plasmodium* at the level of the entire genome. In this manner, complex processes involving the interaction of multiple genes, such as stage specific differentiation or response to drug pressure, can be dissected. Such global analysis has, in fact, been made possible with the development of high-throughput techniques in other systems; these include differential display [7], micro-arrays [8], and serial analysis of gene expression or SAGE [9]. Two of these technologies have been applied in malarial systems. Differential display in drug resistant strains of *P. falciparum* identified two genes specifically induced under chloroquine pressure [10], while differences in mRNA expression between sexual and asexual blood stages were recently characterized by shot-gun micro-arrays [11]. In the present study, the SAGE technique for *P. falciparum* has been developed and optimized.

Both SAGE and microarrays are extremely powerful techniques with which to characterize differential gene expression on a global scale. 'Closed' profiling platforms such as microarrays provide rapid means of screening large numbers of experimental samples; however, the expression data is limited to a pre-determined or known set of genes being screened. On the other hand, 'open' platforms such as SAGE can identify expressed genes that have not yet been cloned, genes that are partially sequenced, or novel genes that cannot be identified from sequence information alone. As such SAGE is particularly well suited for *Plasmodium* species whose genomes are not completely annotated. Moreover, by qualitatively and quantitatively analyzing thousands of transcripts from a given population at the same time, SAGE achieves a greater depth of coverage and readily detects low abundant transcripts. Furthermore the technique may prove useful for identifying the function of many of the unknown genes catalogued by the Genome project as well as assigning ORFs to previously uncharacterized genome sequence reads.

SAGE is based on three experimentally confirmed principles [9]. First, a short (10bp) tag from a defined position within a transcript can uniquely identify a gene. This is reasoned by the fact that the maximum number of possible tag sequences, assuming a random nucleotide distribution, is far greater than the number of estimated genes in most organisms. Second, concatenation of several tags into a single molecule allows for efficient sequencing and acquisition of data. And third, expression patterns of induced or repressed genes are accurately represented by the abundance of their corresponding tags.

A brief description of the generation and isolation of SAGE tags follows (see Fig. 1). A more detailed account of SAGE can also be obtained from Velculescu et al. [9]. cDNA from the population of interest is digested with an anchoring enzyme (AE) that is expected to cleave most transcripts at least once. The AE defines the position of each SAGE tag along a transcript. Linker molecules (40 bp) are subsequently attached to the digested cDNA. These molecules contain recognition sites for a type IIS restriction enzyme that will bind the linker and cleave 12–16 bp away from its binding site to release a short cDNA tag (SAGE tag). The released molecules (consisting of a 40 bp linker and a ~12 bp cDNA tag) are ligated to each other, forming ~102 bp structures containing ditags (two tags linked tail to tail). These 102 bp fragments are amplified using primers that bind to the 40 bp linkers. Purified 102 bp DNA is cleaved with the AE to release 22–24 bp ditags, which are ligated into long concatemers for cloning into a plasmid vector. The plasmids containing ditag concatemers are sequenced, yielding quantitative data on abundance of each SAGE tag. The transcript, from which each SAGE tag was derived, is identified through analysis of sequence databases using software tools.

SAGE has been successfully applied in a number of different systems. For example, it has been used to: (a) characterize the entire repertoire of expressed transcripts in yeast [12]; (b) identify p53 regulated genes [13,14]; (c) compare differential gene expression between normal and cancerous human cells [15–18]; and (d) profile gene expression in rice seedlings [19]. In summary, SAGE lends itself as an extremely efficient tool for qualitative and quantitative monitoring of global gene expression. The high level of accuracy and sensitivity, as well as the depth of coverage achieved by SAGE accounts for its comparative advantage over other methods of transcript profiling (http://www.genzyme.com).

Here, the application of this technique in *P. falciparum* is reported. It was possible to generate a preliminary SAGE library of 6880 tags from the asexual blood stages of 3D7 strain parasites. To the authors' knowledge, this is the first use of SAGE for profiling

malarial gene expression. More importantly, it was demonstrated that SAGE is feasible in an organism whose genome is A–T rich. For instance, a sequence as short as 10 bp was found to be sufficient for uniquely identifying parasite genes encoded in both the nuclear and mitochondrial genomes. Moreover the high A–T content did not preclude the use of restriction enzymes that have been employed in other systems for isolating tags and hence generating representative SAGE libraries. Finally, a few modifications to the DNA extraction and cloning steps of the SAGE protocol proved useful for bypassing specific problems presented by the A–T richness of *Plasmodium* sequence.

2. Methods

2.1. Primers and linkers

Biotinylated oligo(dT)20 was obtained from Gibco BRL and the remaining oligonucleotides were obtained from Integrated DNA Technologies. All primers were SDS-PAGE purified. SAGE linker 1 was formed by hybridizing oligonucleotide 1B (5′-TCCCTATTAAGCCTAGTTGTACTGCACCAGCAAATCC-3′) to oligonucleotide 1A (5′-TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATAGGGACATG-3′). SAGE linker 2 was formed by hybridizing oligonucleotide 2B (5′-TCCCCGTACATCGTTAGAAGCTTGAATTCGAGCAG-3′) to oligonucleotide 2A (5′-TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGGGGACATG-3′). Oligonucleotide 1B and 2B included a 3′ C7 amino modification and were phosphorylated at their 5′ end. SAGE linkers 1 and 2 were self ligated and run on a 12% polyacrylamide gel to determine phosphorylation efficiency. Only linker pairs that self-ligated to form di-linkers at an efficiency of 70% or greater were used in subsequent steps. SAGE linkers 1 and 2 each contain an overhang for the AE. NlaIII (NEB) which recognizes the sequence 5′ CATG 3′ was used as the AE.

Figure 1.

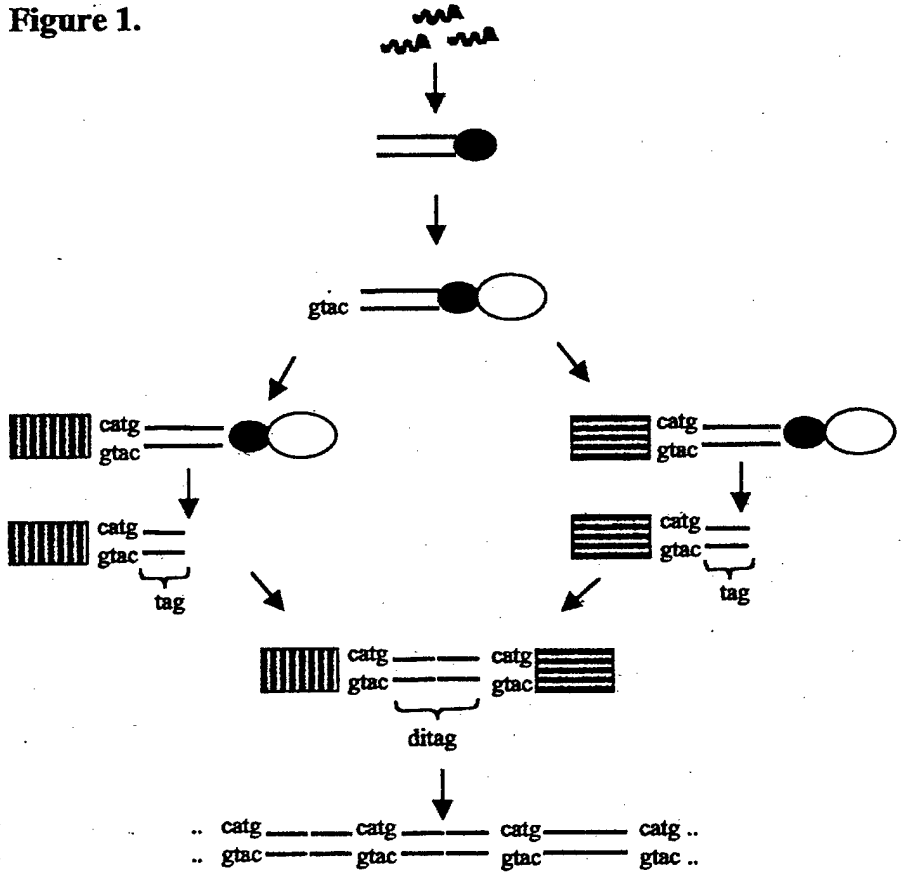


Fig. 1. Schematic illustration of the SAGE technique. (1) 3D7 mRNA; (2) is transcribed into double stranded cDNA with a biotinylated oligo(dT) primer (black oval); (3) cDNA is digested with an AE and bound to streptavidin beads (white oval) to isolate the 3′ most AE site of each transcript; (4) samples are divided and ligated to one of two linkers (hatched boxes); (5) samples are digested with a typeIIS enzyme to release linkers attached to a 10 bp SAGE tag; (6) released tags are blunt-ended and ligated together to form a 102 bp molecule, which is PCR amplified. 22mer ditags are isolated (7) Ditags are ligated to form concatemers, which are cloned and sequenced.

Linkers 1 and 2 also contain sequences recognized by a type IIS restriction enzyme (called the tagging enzyme) and a priming site for PCR. BsmFI (NEB) which recognizes the sequence 5′ GGGAC 3′ served as the tagging enzyme. PCR primer sequences for linker 1 and 2 were 5′-GGATTTGCTGGTGCAGTACA-3′ and 5′-CTGCTCGAATTCAAGCTTCT-3′, respectively.

2.2. Parasite cultures

P. falciparum clone 3D7 (kindly provided by Dr Dan Carucci, Naval Medical Research Center) was maintained in continuous culture as described by Trager and Jensen [20], with some modifications. Briefly, cultures were grown in RPMI (supplemented with 0.5% Albumax I, 24 mM sodium bicarbonate, 11 mM glucose, 12 mM TES sodium salt, 1 mM pyruvate, 2 mM glutamine, 0.04 mM hypoxanthine, 0.0005% Gentamycin) at a 5% hematocrit in a 5% carbon dioxide, 1% oxygen and balanced nitrogen environment. Cultures were placed on a shaking platform to minimize multiple invasion of red blood cells (rbcs) by merozoites. Media and gas were replaced every day and the percentage of parasitized rbcs (parasitemia) was determined by thin blood smears. Cultures at parasitemias of approximately 12%, where the majority consisted of trophozoite forms (8% trophozoites, 2% rings, 2% schizonts), were harvested for isolation of total RNA.

2.3. RNA extraction and cDNA synthesis

Total RNA was extracted immediately using Tri-reagent BD for blood products (Molecular Research Center), and selected twice on oligo(dT) cellulose using the Message Maker reagent assembly (Gibco, BRL) to enrich for mRNA. A total of 1010 parasites typically yield approximately 20 μg of mRNA. The integrity of mRNA samples was checked by gel electrophoresis, RT-PCR and northern blot analysis using probes for PfMDR1 and calmodulin (data not shown).

In separate experiments, total RNA was extracted with either Tri-reagent (Molecular Research Center), or with Tri-reagent following saponin lysis (1% saponin in PBS) of parasitized rbcs, which serves to lyse rbcs while leaving parasites intact. mRNA was purified from total RNA with the Oligotex mRNA kit (Qiagen). Each of the different methods outlined above yielded input mRNA of a high quality, which was successfully used to generate template for amplification of ditags (see subsequent sections).

cDNA was synthesized from 5 μg of mRNA with the cDNA synthesis system (Gibco, BRL) following the manufacturer's recommendations for protocol 1. First strand cDNA synthesis was primed with 2.5 μg of biotinylated oligo(dT)20. The efficiency of oligo(dT) biotinylation was previously verified by measuring percentage binding to streptavidin beads (Dynal). The quality of double stranded cDNA was checked by agarose gel electrophoresis.

2.4. Definition and isolation of cDNA tags

The entire cDNA sample was digested with 100 U of the AE, NlaIII (New England Biolabs-NEB) for 2 h at 37°C, in two reaction volumes of 200 μl each. The digest was extracted with an equal volume of phenol/chloroform/isoamyl alcohol (Sigma) or PCI (25:24:1), precipitated with ethanol [200 μl sample, 3 μl glycogen (Boehringer Mannheim), 100 μl 10 M ammonium acetate (Sigma), 900 μl ethanol] on dry ice for 10 min, and centrifuged at 13 000 rpm for 40 min at 4°C. The pellet was washed once with 70% ethanol and resuspended in 20 μl of LoTE (3 mM Tris–HCl pH 7.5, 0.2 mM EDTA pH7.5-stock solutions from Gibco-BRL).

The 3′ ends of cDNA molecules were isolated through the binding of biotinylated oligo(dT) to paramagnetic streptavidin beads. This process exposes a unique site on each transcript corresponding to its 3′ most AE site. Briefly, the cDNA sample was divided into two fractions (10 μl each). Each fraction was incubated with 1 mg of beads [previously washed with binding/wash solution (5 mM Tris, 0.5 M EDTA, 1 M NaCl)] in 200 μl of binding/wash solution for 30 min at room temperature. The bead-bound cDNA samples were washed twice with 200 μl of binding/wash solution and once with 200 μl of LoTE.

Each fraction was then ligated to either linker 1 or 2 via the AE overhang. Briefly 2 μg of either linker were incubated with bead-bound cDNA in a 40 μl reaction volume at 50°C for 2 min, followed by a 15 min incubation at room temperature. Ten units of T4 DNA ligase together with its supplied buffer (Gibco, BRL) were added to the reaction and placed at 16°C for 2 h. Next the bead-bound cDNA-linker samples were washed four times with 100 μl of binding/wash solution, transferred to a new 1.5 ml tube, and washed once with 100 μl of binding/wash solution and 100 μl of 1 × NEB buffer 4.

Short SAGE tags were released from cDNA molecules by incubation with the tagging enzyme, BsmFI. Briefly the bead-bound cDNA-linker sample was incubated with 2 U of BsmF1 (NEB) for 1.5 h at 65°C, in a 100 μl reaction volume. This served to release a 10 bp fragment of cDNA attached to a 40 bp linker molecule. The supernatant was then transferred to a fresh 1.5 ml tube, PCI extracted, ethanol precipitated, washed twice, and resuspended in 11 μl of LoTE.

2.5. Generation and PCR amplification of 102 bp ditag molecules

Released tags were incubated with the Klenow fragment of DNA Polymerase I to produce blunt-ended products, and then ligated to each other to form a ditag-containing molecule. Briefly, each fraction of released tags was blunt-ended by incubation with 5 U of the Klenow fragment of DNA Polymerase I (NEB) for 30 min at 25°C in a 50 μl reaction volume containing 1 × second strand buffer (a component of the cDNA synthesis system; Gibco, BRL), and 0.025 mM each dNTP. The samples were PCI extracted, ethanol precipitated, and washed as above. Pellets were resuspended in 6 μl of LoTE.

A small aliquot of released tags were radiolabeled to assess the quality of manipulations up to this point. Briefly, 1 μl of released tags was incubated with 1 × second strand buffer (cDNA synthesis system; Gibco, BRL), 0.03 mM each of dCTP, dGTP, dTTP, 2.5 U of the Klenow fragment of DNA Polymerase I and 1 μl of (^{32}P dATP as above. The reaction was run on a polyacrylamide gel and exposed wet to autoradiographic film for 20 min at −80°C.

The two fractions of blunt-ended tags were ligated to each other, forming a molecule containing a ditag. Briefly, 2 μl from both fractions were incubated together at 16°C for 16 h, in a 6 μl reaction volume containing 4 U of T4 DNA ligase and the supplied buffer (Gibco, BRL). Two sets of ligation reactions were set up. A control that lacked ligase was also included. The ligation reactions were PCI extracted, ethanol precipitated and resuspended in 30 μl of LoTE.

The ligated products were then amplified by PCR to generate sufficient material from which ditags could be isolated. Briefly, these samples (including the ligase minus control) were diluted 20-fold for use as PCR template. PCR reactions (50 μl reaction composed of 16.6 mM ammonium sulfate, 67 mM Tris pH 8.8, 6.7 mM magnesium chloride, 10 mM β-mercaptoethanol, 6% DMSO, 0.375 mM each dNTP, 350 ng of each SAGE primer, 5 U Taq polymerase (Perkin-Elmer) and 1 μl of template) were set up in Hot Start tubes (Gibco, BRL). Amplification was carried out for 27 cycles of 30 s at 95°C, 1 min at 55°C, and 1 min at 72°C, with initial heat activation for 1 min at 95°C and final extension for 5 min at 72°C. An aliquot of the PCR reaction was resolved on a 12% polyacrylamide gel to check for the presence of a 102 bp product (consisting of a 22 bp ditag flanked on either end by 40 bp linker sequence), expected in ligase plus samples only. An 80 bp product is also generated during PCR (two 40 bp linker molecules ligated together to form a di-linker). To generate sufficient 102 bp product for 22 bp ditag isolation (see next section) multiple PCR reactions (288 or 480) were set up as described above.

The 102 bp product was gel-purified in the following manner. All PCR reactions were pooled, PCI extracted, ethanol precipitated and resuspended in 360 μl of LoTE. The material was run on three 12% polyacrylamide gels and stained with ethidium bromide. Gels were shielded with plexi-glass when visualizing bands with 300 nm UV trans-illumination. The 102 bp band was excised across the width of each gel and divided into three slices. Each gel slice was fragmented by spinning though a 0.5 ml tube, which was pierced with an 18 gauge needle. The sample was collected in a 1.5 ml tube. DNA was eluted from the gels by placing the gel in 300 μl LoTE, and 100 μl 10 M ammonium acetate at 4°C overnight. The samples were heated at 37°C for 2 h, 65°C for 15 min, and gradually cooled to room temperature. Polyacrylamide was removed on SpinX columns (Costar) and the DNA was PCI extracted, precipitated and resuspended in a total volume of 126 μl LoTE.

In separate experiments, SYBR green I stain (Molecular Probes) was used for detecting DNA in polyacrylamide gels. SYBR green is 25 times more sensitive than ethidium bromide, resulting in intense background staining and smearing. This made it difficult to cleanly isolate the 102 bp ditag bands from the 80 bp di-linkers (Fig. 4). Contamination of the 102mer reduces the yield and average size of tag concatemers at subsequent steps [21].

2.6. Isolation of 22 bp ditags and concatenation

Digesting the 102 bp product with the AE resulted in the release of 22 bp ditags. Briefly, the 102 bp gel purified fragment was incubated with 240 U of NlaIII (in two reaction volume of 150 μl each) at 37°C for 2 h. The reaction was PCI extracted, precipitated, resuspended in 32 μl of LoTE, and loaded on three lanes of a 12% polyacrylamide gel. The ditags running at 22–26 bp were excised and eluted as above, except that the heating step at 65°C was omitted. The sample was resuspended in 7 μl of LoTE.

Ditags were concatenated into single molecules, with 5 U of T4 DNA ligase in a total volume of 10 μl, at 16°C for 16 h. Concatenation of tags allows for efficient sequencing of multiple tags from a single clone. The concatemer sample was heated at 65°C for 15 min, chilled on ice for 10 min and loaded on one lane of an 8% polyacrylamide gel. Concatemers resolved as a smear on the gel. Three size fractions (100–400; 400–800; and >800 bp) were excised and gel-purified as for the 102mer. Each size fraction of gel-purified concatemer sample was resuspended in 6 μl of LoTE.

2.7. Cloning and sequence analysis

All manipulations of the pZero-1 plasmid (Invitrogen) were performed using solutions provided by the company to limit the introduction of endonucleases. Two μg of pZero-1 were incubated with 7 U SphI at 37°C for 1.5 h. The digest was extracted, precipitated and resuspended in 60 μl LoTE. The vector sample was diluted 5- and 10-fold, and 1 μl from each dilution was ligated separately to 3 μl of concatemer insert (in a reaction volume of 5 μl) at 16°C for 16 h. Ligation reactions with vector alone were set up as negative controls. A total of 5 U of T4 DNA ligase were used per reaction. The samples were subsequently PCI extracted, ethanol precipitated, washed four times with 70% ethanol and resuspended in 6 μl of LoTE.

Each sample (1 μl) was electroporated into ElectroMax DH10B cells (Gibco, BRL) and plated out on low salt LB plates containing IPTG, zeocin, and X-gal. Colonies were screened directly for inserts by PCR, utilizing the M13 forward and reverse sequences flanking the cloning site as primers. PCR amplification was carried out as before (see Section 2.5), except that 60 ng of each M13 primer and 1 U of Taq polymerase were used in each reaction. An alternate and more rapid method of screening involved lysing bacteria in 50 μl of buffer composed of 50 mM sodium hydroxide, 0.5% SDS, 5 mM EDTA, and 0.025% Bromocresol green at 65°C for 45 min. Samples were mixed with 1 μl of 30% glycerol and run on 1% agarose gel to determine plasmid size. Selected clones were grown in 96 well plate format on a shaking platform at 37°C for 24 h. A total of 50% glycerol stocks of bacterial clones were prepared. Automated sequencing of these was performed with dye terminator chemistry at The Institute of Genomic Research (Rockville, MD) and The Walter Reed Army Institute of Research (Silver Spring, MD) using M13 forward and reverse primers.

Sequence data was analyzed using SAGE software (Genzyme), which identifies cDNA ditag sequence flanked by AE sites in order to extract 14 bp tag counts (4 bp NlaIII site and 10 bp tag sequence), as well as compares experimental tag data to Genbank sequence databases. Basically the software created a database of all potential 14 bp NlaIII tags from Plasmodium sequences registered by the Malaria Genome Consortium in the NCBI Malaria Genetic and Genomics website (www.ncbi.nih.gov/Malaria/) and linked each tag to gene annotations in the NCBI database (as of February 21, 2000). The experimental SAGE library was compared against this data set. When tags matched to sequence reads that have not yet been annotated, a 500–1000 bp sequence surrounding the tag was translated in all reading frames and compared to the entire NCBI protein database using BLASTx.

Tags that did not match the *P. falciparum* NCBI database were searched against a composite assembly of all available genomic *P. falciparum* sequences (as of February 2000), kindly provided by Drs Jessica Kissinger and David Roos (University of Pennsylvania). This database contains the most complete and up-to-date genome sequence from *P. falciparum*, but is not annotated. Hence tags that gave matches were further characterized by the BLASTx function described above. Any 14 bp tags failing to match either database were analyzed as before using only the first 13 bp of tag sequence. The length of an actual tag can vary between 12 and 16 bp since BsmFI does not cut exactly 10 bp away from its recognition site. Other SAGE studies have used 13 bp tags in their analyses [9,19].

3. Results and discussion

In this report, the feasibility of the SAGE methodology as applied to asexual stages of *P. falciparum* is demonstrated. It was possible to successfully generate ditags from parasite cDNA and construct a SAGE library consisting of approximately 6880 tags. Furthermore, it was found that the AE formerly used in other systems (with balanced nucleotide distributions), could effectively define and isolate tags despite the highly rich A–T composition of parasite sequence. In spite of the lower complexity of *P. falciparum* DNA, tags as short as 10 bp were sufficient to uniquely identify parasite transcripts from both nuclear and mitochondrial genomes. This A–T richness may have, however, contributed to decreased ditag yields. Since SAGE is a multi-step process, reduced efficiency of any one single step will impact all those downstream. For example lowered amounts of 22 bp ditag translated into reduced amounts of concatemer insert, which in turn affected cloning efficiency. To help overcome these problems a few modifications were applied to DNA extraction and cloning steps of the established SAGE protocol.

3.1. Occurrence of AE sites

Despite the A–T richness of malarial sequence, it was found that the occurrence of NlaIII sites (5′ CATG 3′) is relatively frequent in parasite DNA; hence NlaIII was chosen as the AE in this system as in others. As mentioned earlier, the AE defines the position of each tag within a transcript and hence should cleave most mRNA molecules at least once, in order to generate truly representative SAGE libraries. In yeast and mammalian genomes, NlaIII cleaves every 256 bp, while most transcripts are much larger. The frequency of NlaIII sites in *P. falciparum* cDNA is lower, around once every 400 bp as calculated from the occurrence of

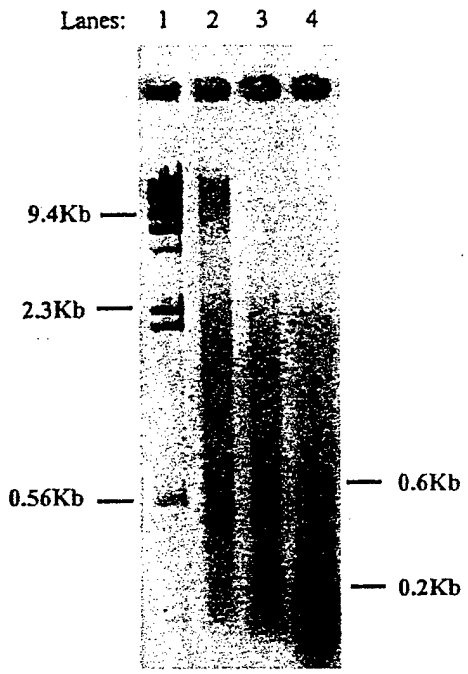


Fig. 2. Size distribution of NlaIII-digested cDNA. Double stranded cDNA from *P. falciparum* was digested with NlaIII and electrophoresed on a 1% agarose gel (lane 3). Undigested cDNA was resolved in lane 2. Lambda DNA-Hind III digest (NEB) (lane 1) and pBR322 MspI digest (NEB) (lane 4) were used as the markers. DNA was stained with ethidium bromide.

CATG in chromosome 2 and 3. After NlaIII digestion, the size distribution of parasite cDNA collapses from between 0.1 and >9.4 kb to between 0.1 and 2.3 kb (see Fig. 2). Since gene density is estimated at one gene every 4.5 kb, NlaIII is still expected to cleave most parasite transcripts. Completion of the Genome Project will reveal those genes which lack NlaIII sites altogether. Another issue relating to the AE is the creation of NlaIII sites by exon splicing. Such sites will be missed in the analysis, since SAGE tags are searched against genomic sequence of *Plasmodium*. EST databases and cDNA sequence will improve the analysis of Plasmodium SAGE tags.

Alternatively, enzymes whose restriction sites occur more frequently in parasite DNA could serve as the AE. However, both NdeI and VspI were tested as the anchoring enzyme in parallel experiments (enzymes that recognize the A–T rich sequences 5′-CATATG-3′ and 5′-ATTAAT-3′, respectively), and discovered that these were not appropriate. In both experiments, sufficient 102mer was produced, but detectable amounts of 22 bp ditag were not released upon digestion with either enzyme. To understand the basis of this result, the 102 bp fragments were cloned into T–A vectors and the presence of ditags subsequently checked by sequence analysis. Approximately half of the 102 bp fragments contained bona fide ditags, while half consisted of two linker molecules each flanking 22–28 bp stretches of As or Ts that lacked the AE site. It was postulated that these aberrant 102mers may represent oligo(dT) that was carried over from the first strand cDNA synthesis reaction, and subsequently annealed to the short 3′–5′ A–T overhangs present on linkers designed to ligate to NdeI- or VspI-digested cDNA. From these data, both the frequency of restriction sites as well as the overhang need to taken into account, and the effectiveness of new anchoring enzymes will have to be determined empirically.

3.2. Generation of 102 bp PCR product and 22 bp ditags

A key step in the SAGE methodology is the generation of 102 bp PCR product since its yield determines the amount of 22 bp ditag recovered. An increased yield of 22 bp ditag in turn enhances cloning and sequencing efficiency downstream. It was demonstrated that sufficient amounts of 102 bp product can be generated from parasite cDNA.

To monitor the quality of manipulations leading up to 102 bp formation, cDNA tags released by BsmFI

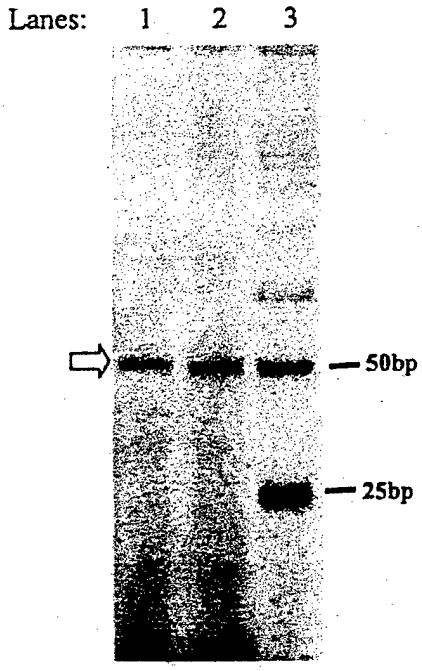


Fig. 3. Blunt-ended cDNA tags. Released cDNA tags from either one of two fractions (one fraction was ligated to linker 1 and the other to linker 2) were incubated with the Klenow fragment of DNA polymerase I and radiolabeled dATP to produce blunt-ended molecules. The blunt-ended tags from fraction 1 (lane 1) and fraction 2 (lane 2) are visible at ~50 bp (40 bp linker + 10 bp cDNA tag) (solid arrow). A 25 bp ladder (Gibco) was used as the marker (lane 3). Samples were run on a 12% polyacrylamide gel and visualized by autoradiography.

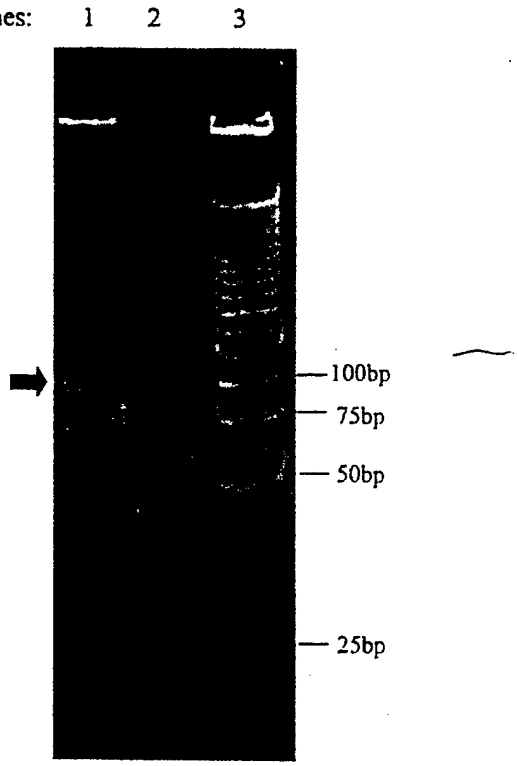


Fig. 4. PCR amplification of ditags. Blunt-ended cDNA tags (attached to linkers) were ligated to each other with T4 DNA ligase and this ligation sample was used as template for 27 cycles of PCR (lane 1). cDNA tags incubated without T4 DNA ligase were also PCR amplified as negative controls (lane 2). The product at ~100 bp in lane 1 corresponds to the amplified ditag containing molecule (see solid arrow). The bands at ~80 bp (lanes1) correspond to di-linker formed by ligation of contaminating 40 bp linker molecules. A 25 bp ladder (Gibco) was used as a marker (lane 3). DNA was run on a 12% polyacrylamide gel and stained with ethidium bromide.

were radiolabeled (Fig. 3). The expected band at approximately 50 bp (40 bp linker attached to a 10 bp cDNA tag) is clearly visible in both fraction 1 and 2. Upon ligation of blunt-ended tags and subsequent PCR amplification, the expected products at 102 bp were obtained (Fig. 4). Bands at 80 bp, corresponding to contaminating di-linker molecules, were also present. Vogelstein and coworkers [22] report a typical yield of 10–20 μg of 102 bp product after gel purification from 96 PCR reactions whereas one obtained 10–20 μg of purified 102 bp product from 288 PCR reactions. This lower yield may be related to the high A–T content of *Plasmodium* sequence. For example, lower AE frequency (as described earlier) could result in fewer sites for generating SAGE tags, and hence reduced amounts of template for PCR amplification of the 102 bp band. In an attempt to optimize PCR, the concentration of dNTP was varied (between 0.0075 and 1.5 mM of each dNTP), and the ratio of dATP/dTTP to dGTP/dCTP (1.5 mM dATP/dTTP: 1.5 mM dCTP/dGTP; 3 mM:1 mM; and 0.3 mM:0.1 mM). Varying the relative ratios of dNTPs in this manner was previously shown to improve PCR amplification of A–T rich mitochondrial sequences [23]. Platinum Taq polymerase was also tested (Gibco, BRL). These modifications did not improve PCR amplification of 102 bp product; instead it was observed that dNTP concentrations above 0.375 mM each were inhibitory to PCR. Hence, in some experiments, 20 μg of 102 bp product was obtained by increasing the number of scale-up PCR reactions by approximately 2-fold.

NlaIII digestion of the 102 bp molecule released 22 bp ditags as expected (Fig. 5). The quantity and quality of these ditags are crucial for determining downstream concatemerization and cloning efficiency. Other studies report ditag yields of several hundred nanograms (500–1000 ng) [21]. One has consistently obtained 100–300 ng of 22 bp ditags, despite increases in PCR scale up. This indicates that, for SAGE in *P. falciparum*, the same amount of 102 bp product yields significantly lower quantities of 22 bp ditags compared to other systems. This reduced yield may be due to the fact that A–T rich SAGE tags are predicted to have lower

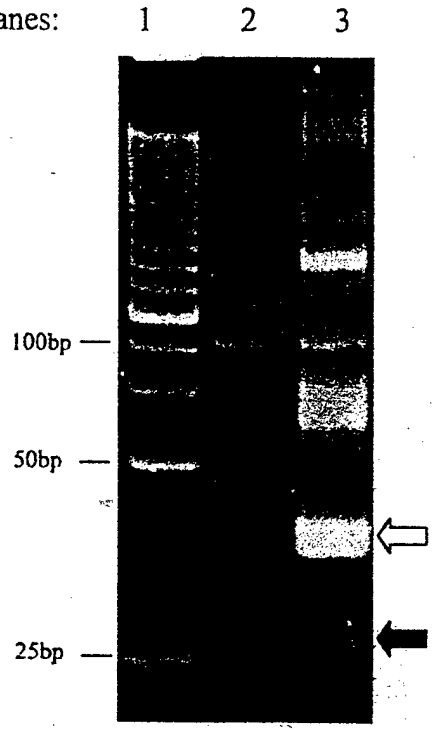


Fig. 5. Isolation of 22 bp ditags. The 102 bp PCR amplified product was gel purified and digested with NlaIII to cleave off 40 bp linkers from both ends (open arrow, lane 3) and to release the 22 bp ditag (solid arrow, lane 3). The band at ~75 bp corresponds to partially digested 102 bp products. Undigested gel purified 102 bp product is shown in lane 2. A 25 bp ladder (Gibco) was used as a marker (lane 1). DNA was run on a 12% polyacrylamide gel and stained with ethidium bromide.

Table 1
Summary of 3D7 SAGE library analysis

	3D7 parasite population
Total number of clones	575
Total number of tags	6880
Total number of tags after excluding linker-derived tags	6702
Average number of tags/clone	12
Percentage of duplicate ditags[a]	3
Total number of genes	4146

[a] Duplicate ditags include ditag sequences that are observed more than once. The percentage of duplicate ditags is calculated by multiplying the number of duplicate ditags by 2 and dividing by the total number of tags.

melting temperatures, which in turn might result in the loss of 22 bp ditags during gel electrophoresis and extraction. Addition of salt, gradual cooling of the DNA sample to room temperature, and minimization of UV exposure during extraction of both the 102mer and 22 bp ditags marginally improved the yield as assessed by cloning efficiency, since the percentage of clones containing concatemer inserts increased by 10% (see next section).

The second parameter is the purity of ditags in the concatemer reaction. After PCR amplification and NlaIII digestion of the 102 bp PCR product, the resulting products contained large amounts of 80 bp di-linker and 40 bp linker respectively. Although the 102 and 22 bp ditags are gel-purified, excess linker material could run aberrantly on the gel and serve to poison the concatemerization reaction, by preventing the extension and cloning of concatemers. Excessive contamination with 80 bp di-linkers was ruled out by gel analysis of a small aliquot of the purified 102 bp product (Fig. 5), which showed by ethidium bromide staining that the sample contained $<1\%$ contamination. Surprisingly, upon digestion of the 102 bp fragment ($>99\%$ pure), the relative ratio of 22 bp ditag to 40 bp linkers was approximately 1:8 in the present study while this ratio should be approximately 1:2. Hence stoichiometric amounts of ditag insert were not recovered in the present study. Perhaps, due to their high A–T content, the 22 bp molecules dissociate and are lost during gel electrophoresis.

3.3. Cloning

Ditags were successfully concatenated and cloned into the pZero vector. Each clone contained an average of 12 tags (see Table 1), a number similar to those obtained by several other investigators [24,25]. However, cloning efficiency was compromised by lower ditag yields in preceding steps. Several modifications were applied to the cloning protocol accordingly.

Upon transformation of *Escherichia coli* with ligated plasmids, pZero utilizes its ccdB-LacZ fusion gene as a lethal selection against colonies containing no insert (Invitrogen); however, in the authors' hands, this selection was leaky. Hence, decreased insert concentrations resulted in an extremely low frequency of insert-containing colonies (4%). This problem was compensated for by reducing the concentration of pZero in the ligation by 5- and 10-fold. This served to increase the frequency of clones with inserts. One routinely obtains frequencies between 30 and 47%, with the higher frequencies relating to the use of high salt elution during preparation of 102 and 22 bp molecules.

In order to bypass the need for screening clones, the transformations was plated on X-gal plates, exploiting the LacZ marker in a double selection. Of the clones grown in the absence of X-gal, 31% were positive for insert; on the other hand, 84% of all white colonies selected from X-gal plates contained insert. Double selection of clones may prove useful when establishing SAGE libraries in systems with limited amounts of RNA [25–27].

3.4. Sequence analysis

A SAGE library of approximately 6880 tags has been generated from 3D7 strain parasites. This library is currently being expanded to provide a more comprehensive expression profile, which will be presented in Patankar et al. (in preparation). A preliminary analysis of sequence data from the current library is presented in Table 1. A total of 4146 different genes were represented in the 3D7 library. Of these, 1047 genes were represented by tags at an abundance level of 2 or greater (some tags at an abundance of 1 may be produced by sequencing error). The percentage of linker derived tags, i.e. tags corresponding to linker sequence, as well as the percentage of duplicate ditags (repeated ditags) were both only 3%. The percentage of duplicate ditags provides a measure of biased PCR. In other SAGE studies, this percentage ranges from between 4% to as much as 10% (http://www.sagenet.org).

To determine whether 10 bp SAGE tags could uniquely identify parasite genes, blast analysis of the 256 most abundant tags (tags at abundance level of 4 or greater) were conducted: 66% of these tags matched to unique sites in the Plasmodium genome, 20% of the tags did not match registered sequence in either database, and 14% matched to more than one locus. In other systems, there have also been several instances where two or more genes share the same tag, i.e. some SAGE tags match to more than one locus in the sequence database. Northern blot analysis of tags in all three classes is underway and should help resolve whether tags that match multiple genes indeed represent multiple transcripts.

Examples of highly abundant SAGE tags are listed in Table 2. These were derived from both the nuclear genome as well as the 6 kb mitochondrial element. Tags in the latter group map to intergenic regions of the mitochondrial mRNAs, where small (40–190 nt) highly fragmented rRNA molecules are encoded on both DNA strands [28,29]. The 6 kb element is polycistronically transcribed, and transcripts containing adjacent mRNA and rRNA sequences have been found [30]. Hence it is unclear whether the SAGE tags are indeed derived from rRNA molecules or from precursor intermediates of polycistronic transcription in the mitochondria. Interestingly thioredoxin, a nuclear encoded transcript is also expressed at high levels, consistent with the abundant expression of genes involved in maintaining mitochondrial physiology and function. Tags corresponding to parasite specific surface proteins that are required for erythrocyte invasion such as MSP-3 (merezoite surface protein 3) SERA (serine repeat antigen), and Rhop H3 (rhoptry protein) were also abundant. Finally several unknown genes and hypothetical ORFs were also highly expressed. Hence SAGE will prove useful for assigning ORFs to previously uncharacterized sequence reads generated by the Genome project.

Additionally it was found that 11 of the 256 most abundant transcripts were represented by more than one 10 bp tag sequence; nine genes were represented by two different tags each, and two genes by three different tags. This phenomenon could result from a partial digest of parasite cDNA, thereby generating many tags at sites other than the 3′ most NlaIII site for a given gene. However, in such a scenario, the 3′ most NlaIII site of a transcript might be expected to generate the most abundant SAGE tag; in the present study, the most abundant tag for a given gene was not always the one located at the most 3′ position. Hence, it was postulated that the high A–U content of *P. falciparum* RNA permits internal priming by oligo(dT) during the cDNA synthesis step of the SAGE protocol, resulting in the generation of more than one tag from a single gene. In fact similar internal priming by oligo(dT) at poly(A) stretches within mRNA transcripts was observed in other SAGE studies [31] as well as during RT-PCR amplification of *P. falciparum* genes corresponding to ABRA (acidic basic repeat antigen), SERA and elongation factor, eF-1a in the laboratory (data not shown). Multiple tags that match a single gene have been observed in other systems; the abundance of such a transcript was calculated by adding all multiple tag counts [12,32].

Interestingly, internal priming of cDNA at poly(A) tracts within genes confers some advantages to the SAGE procedure in *P. falciparum*. For example, it is unclear whether the entire pool of mitochondrial transcripts in the malarial parasite is polyadenylated; transcriptional mapping of mitochondrial mRNA and rRNA molecules has shown that these possess very short (6–20 bp) or non-existent poly (A) tails [33,34]. However tags mapping to the 6 kb mitochondrial element were found at high abundance as mentioned earlier. Hence the apparent lack of long poly (A) tails on mitochondrial transcripts did not exclude them from representation in the SAGE library. Binding of internal poly (A) tracts within parasite transcripts to oligo (dT) columns during the mRNA selection steps of SAGE as well as internal priming during the cDNA synthesis step allows the inclusion of differentially polyadenylated RNAs in the analysis.

In conclusion, it was demonstrated that *P. falciparum* is amenable to the SAGE technique, despite its low genome complexity. In conjunction with other methods of high-throughput transcriptional analysis such as micro-arrays [8], SAGE should yield valuable information about the fundamental biology and virulence mechanisms of an important human pathogen.

Table 2
Highly expressed genes[a]

Tag sequence	Database match	Abundance
TCAGGCGTTA	Mitochondrial 6 kb transcript	1.30
GAGCAAGCAG	No match	0.58
ATTTGAAGCA	Rhop H3	0.42
CTCAGCCGCC	Mitochondrial 6 kb transcript	0.39
GTAGTTGACA	Mitochondrial 6 kb transcript	0.36
CGAGGAAAAA	SERA	0.27
AACGACAAGA	Pfg 27/25	0.25
TACAGCTGCT	MSP-3 (merezoite surface protein 3)	0.20
GGGAAAGCGA	Hypothetical ORF	0.19
GGCACAACTA	Thioredoxin	0.16
GGATATAAAA	Unknown protein	0.16

[a] Examples of the highly abundant SAGE tags from the 3D7 control library and their corresponding genes are listed. Tag sequence represents the 10 bp SAGE tag sequence adjacent to the NlaIII site. Abundance is listed as a percentage of all 6702 tags in the SAGE library.

Acknowledgements

We are indebted to Dr Clarence Wang (Genzyme Molecular Oncology) for providing the SAGE software and for help with data analysis. We also thank Drs J. Kissinger and D. Roos (University of Pennsylvania; droos@sas.upenn.edu) for providing access to assembled *P. falciparum* genomic sequences. We acknowledge the invaluable data provided by the Malaria Genome Consortium: Sequence data for *P. falciparum* chromosome (1,3,4,5,6,7,8,9,13) was obtained

from The Sanger Centre website at http://www.sanger.ac.uk/Projects/P_falciparum/. Sequencing of *P. falciparum* chromosom (1,3,4,5,6,7,8,9,13) was accomplished as part of the Malaria Genome Project with support by The Wellcome Trust. Sequence data for *P. falciparum* chromosome 12 was obtained from the Stanford DNA Sequencing and Technology Center website at http://www-sequence.stanford.edu/group/malaria. Sequencing of *P. falciparum* chromosome 12 was accomplished as part of the Malaria Genome Project with support by the Burroughs Wellcome Fund. Preliminary sequence data for *P. falciparum* chromosome (2,10,11,14) was obtained from The Institute for Genomic Research website (www.tigr.org). Sequencing of chromosome (2,10,11,14) was part of the International Malaria Genome Sequencing Project and was supported by awards from the Burroughs Wellcome Fund and the U.S. Department of Defense. The Chromosome 2 Sequencing Project was a collaborative effort by The Institute of Genomic Research (TIGR) and the Naval Medical Research Center (NMRC). We also thank Drs. Connie Chow and Sarah Volkman for their insightful comments about this manuscript. This work was supported by the Burroughs Wellcome Fund.

References

[1] WHO. The World Health Report. Conquering, Suffering, Enriching Humanity. Geneva: WHO publishers 1997.

[2] Butler D. Funding assured for international malaria sequencing project [news]. Nature 1997;388:701.

[3] Craig AG, Waters AP, Ridley RG. Malaria genome project task force: a post-genomic agenda for functional analysis [news]. Parasitol Today 1999;15:211–4.

[4] O'Brien C. Malaria genome project gets a funding boost [news]. Mol Med Today 1997;3:3.

[5] Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T, et al. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum* [see comments]. Nature 1999;400:532–8.

[6] Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C, et al. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum [published erratum appears in Science 1998 Dec 4;282(5395):1827]. Science 1998;282:1126–32.

[7] Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction [see comments]. Science 1992;257:967–71.

[8] Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–62.

[9] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression [see comments]. Science 1995;270:484–7.

[10] Thelu J, Burnod J, Bracchi V, Ambroise-Thomas P. Identification of differentially transcribed RNA and DNA helicase-related genes of *Plasmodium falciparum*. DNA Cell Biol 1994;13:1109–15.

[11] Hayward RE, DeRisi JL, Alfadhli S, Kaslow DC, Brown PO, Rathod PK. Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria [In Process Citation]. Mol Microbiol 2000;35:6–14.

[12] Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Jr, Hieter P, Vogelstein B, Kinzler KW. Characterization of the yeast transcriptome. Cell 1997a;88:243–51.

[13] Madden SL, Galella EA, Zhu J, Bertelsen AH, Beaudry GA. SAGE transcript profiles for p53-dependent growth regulation. Oncogene 1997;15:1079–85.

[14] Polyak K, Xia Y, Zweier JL, Kinzler KW, Vogelstein B. A model for p53-induced apoptosis [see comments]. Nature 1997;389:300–5.

[15] Hibi K, Liu Q, Beaudry GA, Madden SL, Westra WH, Wehage SL, Yang SC, Heitmiller RF, Bertelsen AH, Sidransky D, et al. Serial analysis of gene expression in non-small cell lung cancer. Cancer Res 1998;58:5690–4.

[16] Hibi K, Westra WH, Borges M, Goodman S, Sidransky D, Jen J. PGP9.5 as a candidate tumor marker for non-small-cell lung cancer. Am J Pathol 1999;155:711–5.

[17] Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K, et al. A public database for gene expression in human cancers. Cancer Res 1999;59:5403–7.

[18] Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW. Gene expression profiles in normal and cancer cells. Science 1997;276:1268–72.

[19] Matsumura H, Nirasawa S, Terauchi R. Technical advance: transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE). Plant J 1999;20:719–26.

[20] Trager W, Jensen JB. Human malaria parasites in continuous culture. Science 1976;193:673–5.

[21] Powell J. Enhanced concatemer cloning — a modification to the SAGE (Serial Analysis of Gene Expression) technique. Nucleic Acids Res 1998;26:3445–6.

[22] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of Gene expression: detailed protocol (version 1.0c). Available from Johns Hopkins Oncology Center and Howard Hughes Medical Institute, 424 North Bond Street, Baltimore, MD 21231, 1997.

[23] Rondan Duenas JC, Panzetta-Dutari GM, Gardenal CN. Specific requirements for PCR amplification of long mitochondrial A + T-rich DNA. Biotechniques 1999;27:258–60.

[24] Ryo A, Suzuki Y, Ichiyama K, Wakatsuki T, Kondoh N, Hada A, Yamamoto M, Yamamoto N. Serial analysis of gene expression in HIV-1-infected T cell lines. FEBS Lett 1999;462:182–6.

[25] Datson NA, van der Perk-de Jong J, van den Berg MP, de Kloet ER, Vreugdenhil E. MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acids Res 1999;27:1300–7.

[26] Peters DG, Kassam AB, Yonas H, O'Hare EH, Ferrell RE, Brufsky AM. Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. Nucleic Acids Res 1999;27:e39.

[27] Virlon B, Cheval L, Buhler JM, Billon E, Doucet A, Elalouf JM. Serial microanalysis of renal transcriptomes. Proc Natl Acad Sci USA 1999;96:15286–91.

[28] Feagin JE, Werner E, Gardner MJ, Williamson DH, Wilson RJ. Homologies between the contiguous and fragmented rRNAs of the two *Plasmodium falciparum* extrachromosomal DNAs are limited to core sequences. Nucleic Acids Res 1992;20:879–87.

[29] Feagin JE, Mericle BL, Werner E, Morris M. Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6 kb element. Nucleic Acids Res 1997;25:438–46.

[30] Ji YE, Mericle BL, Rehkopf DH, Anderson JD, Feagin JE. The *Plasmodium falciparum* 6 kb element is polycistronically transcribed. Mol Biochem Parasitol 1996;81:211–23.

[31] de Waard V, van den Berg BM, Veken J, Schultz-Heienbrok R, Pannekoek H, van Zonneveld AJ. Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus. Gene 1999;226:1–8.

[32] Kal AJ, van Zonneveld AJ, Benes V, van den Berg M, Koerkamp MG, Albermann K, Strack N, Ruijter JM, Richter A, Dujon B, et al. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Mol Biol Cell 1999;10:1859–72.

[33] Rehkopf DH, Gillespie DE, Harrell MI, Feagin JE. Transcriptional mapping and RNA processing of the *Plasmodium falciparum* mitochondrial mRNAs. Mol Biochem Parasitol 2000;105:91–103.

[34] Gillespie DE, Salazar NA, Rehkopf DH, Feagin JE. The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum* have short A tails. Nucleic Acids Res 1999;27:2416–22.

Molecular and Biochemical Parasitology, 2001; in press.

Deletion Analysis of the 5' Flanking Sequence of the *Plasmodium gallinaceum* Sexual Stage Specific Gene *pgs28*[#] Suggests a Bipartite Arrangement of *cis*-Control Elements

Wilfred F. Mbacham[1,2], Connie S. Chow[1], Johanna Daily, Linnie M. Golightly[3], and Dyann F. Wirth*.

Department of Immunology and Infectious Diseases, Harvard School of Public Health, 665 Huntington Avenue, Boston MA 02115

[1] Equal contribution.

[2] Present address: The Biotechnology Center, Faculty of Science. University of Yaounde, Box 812, Yaounde, Cameroon.

[3] Present address: Department of Medicine, Weill Medical College of Cornell University, New York, NY 10021.

* Corresponding author. Tel: 617-432-1563; fax: 617-432-4766; email: dfwirth@hsph.harvard.edu

[#]GenBank accession no. AY007246

Keywords: promoter, transcription, gene expression, ookinete, luciferase, *cis*-control elements.

Abbreviations: pBS, pBluescript vector; EGF, epidermal growth factor; spp., species; GUS, glucouronidase

The malaria parasite undergoes a complex developmental process through its life cycle. This includes an asexual intraerythrocytic cycle in the vertebrate host, and a sexual cycle that commences with gametogenesis in the vertebrate host and subsequent fertilization and maturation in the mosquito vector. Regulation at the transcriptional and post-transcriptional levels is no doubt important for the temporal expression of genes required at each stage of development. Present understanding of the *cis*-elements important for transcriptional control in *Plasmodium* is severely restricted. Sequence analysis of 5' flanking regions of *Plasmodium* genes reveal the presence of sequences with homology to known eukaryotic control elements, for example see [1, 2]; however, the functional significance of these sequences in *Plasmodium* has not been demonstrated. The intergenic region in *Plasmodium spp.* is particularly AT-rich, even within the context of the AT-biased (~80%) genome [3], such that even the identification of TATA-like elements, and assays to determine their utility and importance, becomes difficult. A growing but limited number of functional analyses of promoter regions of *Plasmodium* genes have been published, many of which shed light on regions that are necessary for efficient expression [2, 4]. However, only a few studies to date have identified specific sequences, short of transcriptional start sites, that appear to be important for gene expression [4-6]. Due to the small numbers, and the fact that these genes are expressed at different stages in the parasite life cycle, no consensus or common sequences could be established. Clearly, much more can be learned about aspects of basal transcription as well as stage-specific control of gene expression in the malaria parasite.

Pgs28 is expressed abundantly on the surface of mosquito stages of the avian parasite, *Plasmodium gallinaceum*. Pgs28 belongs to the family of Pxs proteins, which includes the *P. berghei* homolog Pbs21 and the *P. falciparum* homolog Pfs25. These proteins contain a series of EGF-like domains that may serve a function in cell signalling or in adhesion [7, 8]. Pgs28, Pfs25 and Pbs21 had been identified as targets for transmission blocking antibodies [9-12]. Transcripts of *pbs21* had been observed in female gametocytes and gametes, as well as zygotes and ookinetes, and the *pfs25* promoter appears to be specifically active in mosquito stage parasites, supporting the notion that the genes encoding this protein family are activated specifically during the sexual stages [5, 13, 14]. Since Pbs21 is initially expressed on the surface of zygote stage parasites, additional post-transcriptional control is exerted by the parasite to regulate Pbs21 expression. We are interested in investigating *pgs28* gene expression to further understand transcriptional regulation in *Plasmodium spp.* and as a step towards understanding the control of sexual

development in *P. gallinaceum*. In this report, we present a functional analysis of the 5' flanking region of *pgs28*, using firefly luciferase as a reporter, by which we identified two regions that are required for *pgs28 trans*-gene expression. Furthermore, using Northern analysis, we define the 5' limit of the *pgs28* transcript and demonstrate that *pgs28* transcripts are present during the zygote stage.

The 5' and 3' flanking sequence of *pgs28*, together with an in-frame insertion of the luciferase reporter, were previously cloned into pBS (*pgs28.1LUC)* [15]. In this study, the *pgs28-luc* chimera, containing *pgs28* 5' flanking sequence, the *pgs28-luc* fusion gene, and about 720 bp of 3' flanking sequence, from *pgs28.1LUC* was cloned into the *Hind*III site of pBS to create *BSpgs28-LUC*. The 1871 bp 5' flanking sequence of *pgs28* has been determined and deposited in GenBank. (The sequence and characterization of the 3' region was recently published [16].) Expression from *BSpgs28-LUC* was confirmed by immunofluorescent antibody staining and immuno-electron microscopy [17] and also by luciferase assays performed 24 or 48 hrs post-transfection (see below). High expression levels up to the order of 10^6 light units were obtained, offering a sensitive system for determining changes in expression levels.

In order to determine the sequence requirements for *pgs28* expression, a series of 5' deletion mutants was created either by exonuclease digestion of linearized *BSpgs28-LUC* plasmid, or by PCR mutagenesis. Deletions of 790bp (FP1081), 1131bp (FP740), 1358bp (FP513), 1407 (FP464), 1485bp (FP386), 1538 bp (FP333), 1584bp (FP287), 1631 bp (FP240) and 1905bp (FP+34) from *BSpgs28-LUC* were obtained (Fig. 1A). Additionally, an internal deletion mutant Δ376-316, which lacks the specified sequences, was created by inverse PCR. To assess the contribution of the deleted sequences to *pgs28-luc* transgene expression, these plasmids were transfected into sexual stage parasites as previously described [15]. Luciferase activity was assayed after 24 or 48 hours. To control for transfection efficiency, a second plasmid, *pgs28-GUS*, was co-transfected and luciferase light units normalized to GUS fluorescence units.

Transfection using FP1081 demonstrated that expression of the *pgs28-luciferase* fusion gene did not decrease significantly when the 5' most 790 bp were deleted from the parent plasmid (Fig. 1B). However, luciferase expression from FP740, where an additional 340 bp has been removed, was reduced by more than 40%. A modest decrease in promoter efficiency was observed with the removal of the next 227 bp (FP513). Further deletions of up to 180 bp (FP464, FP386, FP333) did not seem to significantly affect expression when compared to FP513. Interestingly, FP287, containing a deletion of 46 bp 3' of FP333, had less than 5% activity compared to the full-length construct. Furthermore,

the internal deletion mutant Δ376-316 was also severely affected, having only 6.6% activity. As expected, a deletion that encompasses part of the *pgs28* open reading frame (FP+34) abolished luciferase expression. The mutant FP240 had equivalent activity to this construct, suggesting that important elements necessary for transcription and possibly translation have been removed. Taken together, results of the 5' deletion analysis suggest that the minimal sequence necessary for *pgs28* transgene expression consists of the 333 bp upstream of the translational start site. Moreover, a 17 base-pair sequence, TACCATTTGTACAGACAG, between -333 and -316, appears to be crucial, since *pgs28* expression was essentially abrogated in a 5' deletion mutant and an internal deletion mutant that lack these sequences. We suggest that the proximal site corresponds to the basal promoter or initiator element, as indicated in the following section. We also suggest that positive regulatory elements lie between -1081 and -740, and perhaps within -740 and -513. This distal region likely contains an enhancer element(s) that contributes to *pgs28* promoter efficiency. Thus transcriptional elements that control *pgs28* appear to be bipartite, as in eukaryotic promoters and other *Plasmodium* genes that have been analyzed.

We used Northern analysis as a preliminary step to map the transcriptional start site of *pgs28*, and to determine whether the temporal pattern of *pgs28* transcription paralleled that of its murine homolog *pbs21*. RNA was isolated from newly formed zygotes collected after exflagellation, and from gametes. As seen in Fig.2B, an intense signal appeared at a position corresponding to a message of about 1.4 kb in both zygote and gamete when probed with BBm600 (lanes 1 and 2), which extends from -381 in the 5' flanking region to within the *pgs28* coding sequence. A *pgs28* message of 1.5 kb has previously been reported by Duffy and colleagues [11]. Thus, while Pgs28 expression is most abundant on ookinete surfaces, *pgs28* transcript can be seen as early as the zygote stage. This is in agreement with transfection studies in our laboratory using the *BSpgs28* construct, as well as a *pgs28-GFP* fusion, that demonstrated Pgs28 expression on the surface of zygotes [17].

Recently, the polyadenylation signal of *pgs28* was mapped to approximately 425bp downstream of the stop codon, with an estimated poly(dA) tail of at least 20 nucleotides [16]. Given that the coding sequence of *pgs28* is 666bp, the transcription initiation site of *pgs28* would lie approximately between -390 and -290bp. In agreement with this estimation, only the probe BV142, encompassing the sequence from -381 to –240, hybridized to the *pgs28* transcript (Fig. 2B, lane 7), while probes corresponding to sequences further upstream failed to hybridize to *pgs28* mRNA (lanes 3-6). Thus, these studies establish the 5' limit of the *pgs28* transcript at -381 bp upstream from the

translational start site. 5' deletion analysis suggests that the transcriptional start site is likely to be downstream of -333bp. Experiments to determine the precise 5' end of the *pgs28* transcript will resolve this aspect of *pgs28* transcription.

The 5' flanking sequence of *pgs28* had been inspected for homology to other eukaryotic transcriptional regulatory elements. The highly AT-rich region between -1081 and -520, typical of intergenic regions of *Plasmodium spp.*, does not contain sequences that are analogous to known eukaryotic regulatory elements. Two GTAAT sequences, demonstrated to be important for *GBP130* expression [6], can be found in this region. Whether an element associated with an enhancer of an asexual stage gene in *P. falciparum* is important for expression of *pgs28*, a sexual stage specific gene, can only be determined by experimental means.

Sequences downstream of -520 have also been examined. Within this region are two putative TATA elements TAAAAAGAATAA and TATAAATGTTT, centered at -434bp and -360bp respectively from the start codon. Since these sequences can be deleted from the reporter constructs (FP386 and FP333) without drastically affecting expression, they are not likely to be important for *pgs28* expression. This again illustrates that sequence analogy to eukaryotic promoter elements does not necessarily imply functional significance in *Plasmodium* genes. Inspection of the presumed 5' UTR reveals a T-rich stretch, constituting up to 74% of the bases between 130bp and 242bp. A series of five 8-base pair inverse repeat elements (*TTTATTTT*ATTT) could be identified within this sequence. Further examination of this region uncovers 3 direct repeats of 27bp to 29bp in length. Whether these sequences have functions at a post-transcriptional step to enhance *pgs28* expression awaits further experimentation. Recently, transfection studies of *pfs25* promoter constructs into *P. gallinaceum* ookinetes, and mobility shift assays using *P. gallinaceum* ookinete nuclear extracts, suggest that the sequence AAGGAATA, found at -403 to -396 and -483 to -476 from the initiation codon in *pfs25*, interacts with a nuclear factor and is important for expression of *pfs25* [5]. A similar sequence, AAGAATAA, is found at -354 and -347 in *pgs28*, within the putative proximal TATA sequence. Again, the transfection studies reported here suggest that this sequence in *pgs28* can be deleted without severely affecting *pgs28* transgene expression. This suggests that the nuclear factor PAF-1 [5] is not involved in *pgs28* transcription, and/or that it has a stringent sequence requirement that the AAGAATTT sequence in *pgs28* does not satisfy. Even though *pgs28* and *pfs25* belong to the same family, and possess similar expression profiles during the parasite life cycle, they may not necessarily be controlled by the same evolutionarily conserved factors. Nonetheless, given the close evolutionary relationship between *P. gallinaceum* and *P. falciparum*, it would be of great interest to determine

whether the 17 bp upstream sequence in *pgs28* between -333 and -316 would be able to functionally replace the *pfs25* sequence, and vice versa.

Acknowledgment

WFM was supported by a New England Biolabs Foundation Fellowship, Beverly MA, USA and the UNDP/World Bank/WHO Special Program for Research and Training grant # M181/4/1M.353/CER. DFW is supported by RO1 GM61351-01 from the National Institute of Health and DAMD 17-98-1-8003 from the Department of the Army.

References

[1] Su X, Wellems T. Sequence, transcript characterization and polymorphisms of a *Plasmodium falciparum* gene belonging to the heat-shock protein (HSP) 90 family. Gene 1994; 151: 225-30.

[2] Horrocks P, Kilbey B. Physical and functional mapping of the transcriptional start sites of *Plasmodium falciparum* proliferating cell nuclear antigen. Mol Biochem Parasitol 1996; 82: 207-15.

[3] Horrocks P, Dechering K, Lanzer M. Control of gene expression in *Plasmodium falciparum*. Mol Biochem Parasitol 1998; 95: 171-81.

[4] Crabb BS, Cowman AF. Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*. Proc Natl Acad Sci USA 1996; 93: 7289-94.

[5] Dechering KJ, Kaan AM, Mbacham W, Wirth DF, Eling W, Konings RN, Stunnenberg HG. Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. Mol Cell Biol 1999; 19: 967-78.

[6] Horrocks P, Lanzer M. Mutational analysis identifies a five base pair cis-acting sequence essential for *GBP130* promoter activity in *Plasmodium falciparum*. Mol. Biochem. Parasitol 1999; 99: 77-87.

[7] Adini A, Warburg A. Interaction of *Plasmodium gallinaceum* ookinetes and oocysts with extracellular matrix proteins. Parasitol. 1999; 119: 331-6.

[8] Sidén-Kiamos I, Vlachou D, Margos G, Beetsma A, Waters A Sinder R, Louis C. Distinct roles for Pbs21 and Pbs25 in the *in vitro* ookinete to oocyst transformation of *Plasmodium berghei*. J Cell Sci. 2000; 113:3419-26.

[9] Matsuoka H, Kobayashi J, Barker G, Miura K, Chinzei Y, Miyajima S, Ishii A, Sinden R. Induction of anti-malarial transmission blocking immunity with a recombinant ookinete surface antigen of *Plasmodium berghei* produced in silkworm larvae using the baculovirus expression vector system. Vaccine 1996; 14: 120-6.

[10] Kaslow D, Quakyi I, Syin C, Raum M, Keister D, Coligan J, McCutchan T, Miller LH. A vaccine candidate from the sexual stage of human malaria that contains EGF-like domains. Nature 1988; 333: 74-6.

[11] Duffy P, Pimenta P, Kaslow D. Pgs28 belongs to a family of epidermal growth factor-like antigens that are targets of malaria transmission-blocking antibodies. J Exp Med 1993; 177: 505-10.

[12] Paton M, Barker G, Matsuoka H, Ramesar J, Janse C, Waters A, Sinden R. Structure and expression of a post-transcriptionally regulated malaria gene encoding a surface protein from the sexual stages of *Plasmodium berghei*. Mol Biochem Parasitol 1993; 59: 263-75.

[13] Vervenne R, Dirks R, Ramesar J, Waters A, Janse C. Differential expression in blood stages of the gene coding for the 21-kilodalton surface protein of ookinetes of *Plasmodium berghei* as detected by RNA *in situ* hybridisation. Mol Biochem Parasitol 1994; 68: 259-66.

[14] Thompson J, Sinden R. *In situ* detection of Pbs21 mRNA during sexual development of *Plasmodium berghei*. Mol Biochem Parasitol 1994; 68: 189-96.

[15] Goonewardene R, Daily J, Kaslow D, Sullivan TJ, Duffy P, Carter R, Mendis K, Wirth D. Transfection of the malaria parasite and expression of firefly luciferase. Proc Natl Acad Sci USA 1993; 90: 5234-36.

[16] Golightly L, Mbacham W, Daily J, Wirth D. 3' UTR elements enhance expression of Pgs28, an ookinete protein of *Plasmodium gallinaceum*. Mol Biochem Parasitology 2000; 105: 61-70.

[17] Patankar S, Fujioka H, Wirth D. Localization of Pgs28 to the surface of *P. gallinaceum* ookinetes requires the signal sequence and C-terminal hydrophobic domain. Mol Biochem Parasitol 2000; 111:425-35.

Legends

Fig. 1A. Schematic of the *pgs28* 5' flanking sequence.

The 5' flanking sequence of *pgs28* cloned into *BSpgs28-LUC* is shown, together with part of the *pgs28* and *luc* coding sequence. To obtain *BSpgs28-LUC, pgs28.1LUC* [15] was digested with *Hind*III and cloned into similarly digested pBluescript KS+ (Strategene).
The bars at approximately -440 and -360 represent two putative TATA boxes. The hatched box downstream of -240 represent the T-rich sequence with internal repeats.
Positions of the 5' deletion mutants are indicated. The numbers refer to the distance in nucleotides away from the start of the coding region (+1).

To generate the 5' deletion mutants FP1081, FP464, FP513, FP386 and FP+34, *BSpgs28-LUC* was first digested with *SacI* and *SpeI* (New England Biolabs). The linearized plasmid was digested further with exonuclease III/mung bean nuclease essentially as described by the manufacturer (Strategene). *E. coli* (XL-1 Blue) cells were transformed with ligated products and the sizes of the plasmids obtained were determined by agarose gel electrophoresis. FP464 was generated by recloning the filled-in *Nde*I insert from *BSpgs28-LUC* into *Sma*I digested pBS. FP333, FP287 and FP240 were created by PCR mutagenesis, using FP464 as template, and the upstream primers 5'GAATTCCTGCAGCCCTACCATTTGTACAGAC,
5'GAATTCCTGCAGCCCACTAGCTAAAAGAAATATG, and
5'GAATTCCTGCAGCCCATTTTTATTTAATTTTTC respectively. The *Pst*I site is underlined. The downstream primer, 5'CTAGAGGATAGAATGGCGCCG, containing an internal *Sfo*I site (underlined), was used in all cases and was derived from the *luc* coding region. Purified PCR fragments were digested with *Pst*I and *Sfo*I and cloned into similarly cut FP464 vector backbone.

To generate Δ376-316, primers WFM48 5'CCATTTGTTATTGTATATAAAAAAAAAAAC and WFM20R 5'GATCTTCTTAATCTTTGTAAAAATAACTG, which flank the sequences to be deleted, were used to amplify FP513 that had previously been linearized with *Bgl*II, utilizing the *TaqPlus Long* PCR system (Strategene). 30 cycles of PCR reactions were performed in low salt buffer under the following conditions: 94°C for 1 min, 55°C for 1 min and 72°C for 7 mins. PCR products were treated with *Dpn*I at for 30 mins and further treated with 1μl of *Pfu* for 10 cycles and incubation at 37°C for 30 mins. Amplified products were phenol:chloroform extracted and ethanol precipitated, and resuspended. Amplified DNA

containing the deletion was allowed to circularize and transformed into *E. coli* (XL-1 Blue) cells. Sequences of all clones were confirmed by DNA sequencing.

B. Luciferase expression from *pgs28* 5' deletion mutants.

Parasites were transfected with the indicated plasmids and *pgs28-GUS*, and luciferase and GUS activity assessed 24 or 48 hrs post-transfection, as described [16]. The construction of *pgs28-GUS*, containing an in-frame insertion of the *β-glucuronidase* gene (Clontech) within *pgs28*, has been described elsewhere [16]. Normalized relative light units and SD are shown. The indicated activity is the average of 3-8 determinations.

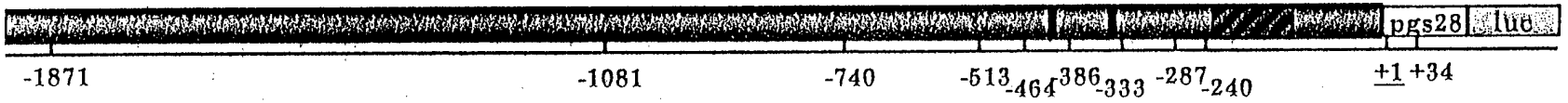
Fig. 2A Positions of DNA probes used to map the 5' end of *pgs28* transcripts.

Probes MS135 (–786 to -651), SN188 (-651 to –463), NB82 (-463 to –381), and BV142(-381 to –240) and BBm600 (-381 to +217) were made by digesting an *XbaI* fragment of *BSpgs28-LUC* containing *pgs28* sequences with *MwoI/SwaI, SwaI/NdeI, NdeI/BglII, BglII/VspI and BglII/BamHI* restriction enzyme pairs, respectively. These resulted in fragments of lengths indicated by the numerals in the designations.

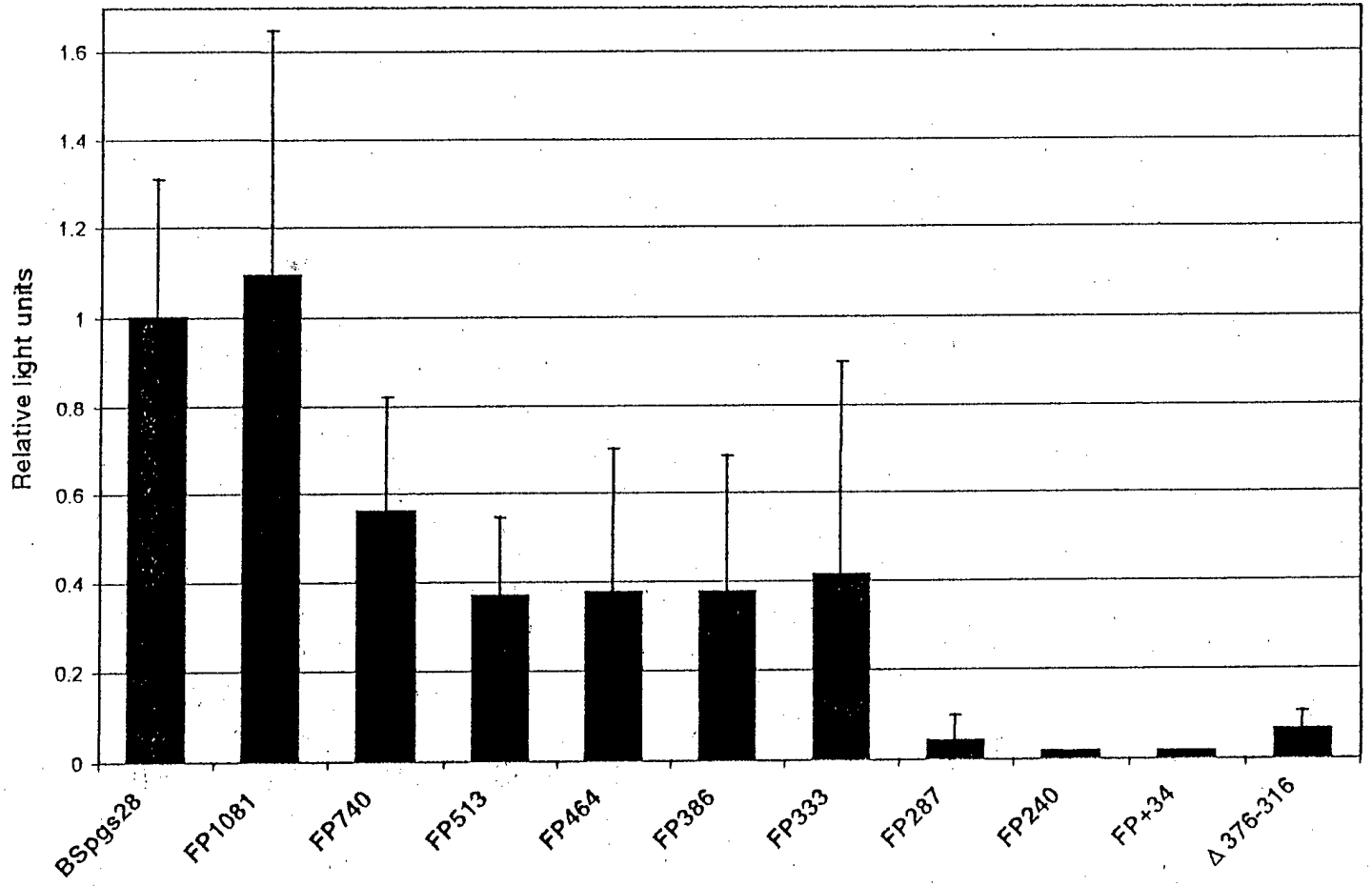
B. Determination of size and the 5' end of *pgs28* mRNA from *P. gallinaceum*

RNA was extracted either from zygotes (lanes 1 and 3) or ookinetes (lanes 2, 4-7), fractionated and Northern blotted using standard procedures. Between four and five micrograms total RNA obtained from $3x10^7$ parasites were included per lane. Blots were probed with the indicated DNA fragments, washed and autoradiographed for 24-48 hours. Lanes 1 and 2, BBm600; lanes 3 and 4, MS135; lane 5, SN188; lane 6, NB82; lane 7, BV142.

A.


pgs28
luc
-1871
-1081
-740
-513
-464
386
-333
-287
-240
+1
+34

B.


Relative light units
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
BSpgs28
FP1081
FP740
FP513
FP464
FP386
FP333
FP287
FP240
FP+34
Δ 376-316

A.

Probes used for Northern analysis:

MS135

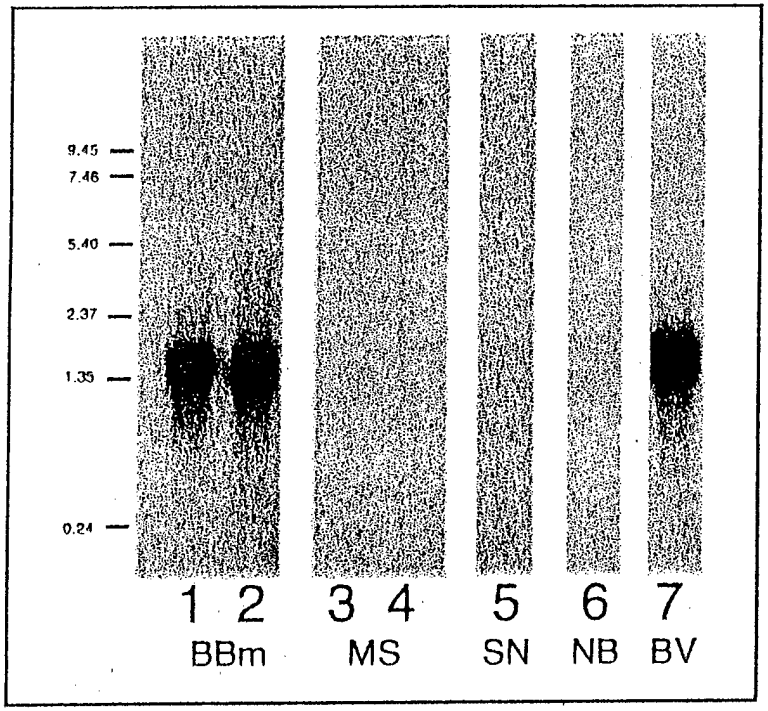
SN188

NB82

BV142

BBm600

B.



Mbacham et al

Serial Analysis of Gene Expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages, as well as novel transcriptional phenomenon in the malarial parasite.

Running title: Transcriptional profiling in the malarial parasite.

Swati Patankar*, Anusha Munasinghe*, Azadeh Shoaibi¶, Leda M. Cummings¶ and Dyann F. Wirth*.

* Department of Immunology and Infectious Diseases, Harvard School of Public Health, Harvard University, Boston MA 02115, USA; ¶The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

S.Patankar and A.Munasinghe made significant contributions to this work.

Corresponding author: D. F. Wirth. Email:dfwirth@hsph.harvard.edu; tel:617-432-1621; fax: 617-432-4766

Abbreviation used: BLAST: Basic Local Alignment Search Tool; SAGE, Serial Analysis of Gene Expression.

Keywords: Serial Analysis of Gene Expression, transcriptional profiling, antisense transcription, *Plasmodium falciparum*.

Abstract

Serial analysis of gene expression (SAGE) was applied to the malarial parasite, *Plasmodium falciparum* in order to characterize the comprehensive transcriptional profile of erythrocytic stages. A SAGE library of approximately 8335 tags representing 4866 different genes was generated from 3D7 strain parasites. BLAST analysis of high abundance SAGE tags identified the major metabolic pathways that are utilized by the organism under normal culture conditions. Furthermore several tags expressed at high abundance (30% of tags matching to unique loci of the 3D7 genome) were derived from previously uncharacterized open reading frames, demonstrating the use of SAGE in genome annotation. The open platform "profiling" nature of SAGE also lead to the important discovery of novel transcriptional phenomenon in the malarial pathogen: a significant number of highly abundant tags that were derived from annotated genes (17%) corresponded to anti-sense transcripts. This SAGE data was validated by two independent means, strand specific RT-PCR and Northern analysis, where anti-sense messages were detected in both asexual and sexual stages. This finding has implications for transcriptional regulation of *Plasmodium* gene expression.

Introduction

Malaria, an infectious disease caused by the protozoan parasite *Plasmodium falciparum*, affects 300-500 million people globally each year (WHO, 1997). Increasing drug-resistance in the parasite and insecticide-resistance in the *Anopheles* vector have exacerbated this substantial public health problem. Against this backdrop, effective strategies to combat the disease require a fundamental knowledge of the basic biology of *Plasmodium* in order to develop new pharmatherapeutics and vaccines that target the parasite.

Most studies of *Plasmodium* biology have been directed at single genes thought to be important for pathogenesis. With the advent of genomic technologies, however, new

approaches to combat the disease, such as identifying entire repertoires of transcripts expressed under different conditions, have now become available. Genomic approaches were initiated with the sequencing of the *P. falciparum* (3D7 strain) genome, a collaborative project, undertaken by the Malaria Genome Consortium that is already close to completion (Butler, 1997; Craig et al., 1999; O'Brien, 1997). Chromosomes 2 and 3 have been fully sequenced (Bowman et al., 1999; Gardner et al., 1998) while eighty to ninety percent of the estimated 6000 open reading frames (ORFs) in the 3D7 genome are now available as raw sequence data. The next challenge is to use this vast amount of data to study the functional relevance of various genes. For example, it is now possible to identify genes that are transcribed in different stages of the parasite's development and also genes that are induced or repressed in response to various stimuli such as immune- or drug- pressure. For this reason, whole genome expression analyses using high-density micro-arrays (Hayward et al., 2000) and serial analysis of gene expression (SAGE) (Munasinghe et al., 2000) have been developed for *P. falciparum*. These new approaches will complement each other to generate data for the *Plasmodium* research community. Genome sequence will expedite the micro-array and SAGE analysis; conversely, open platform profiling techniques such as SAGE will help the Malaria Genome Project with annotation of previously uncharacterized open reading frames (ORFs) and with novel gene discovery.

SAGE provides a sensitive and highly quantitative description of the transcript profile of a given cell type (Velculescu et al., 1995; Velculescu et al., 1997a). The SAGE technology samples short sequence tags (14 bases) from mRNA transcripts in the population of interest. These tags contain sufficient sequence information to identify, by

BLAST analysis, the transcript from which it was derived. The frequency of each tag in the SAGE library is an accurate estimate of the abundance of its corresponding mRNA transcript. Numerous groups have used this technique successfully and described the SAGE protocol in detail (Madden et al., 1997; Matsumura et al., 1999; Polyak et al., 1997; Velculescu et al., 1995; Velculescu et al., 1997a; Virlon et al., 1999).

In this report, we show that SAGE can be used to study gene expression of the asexual stages of *P. falciparum*. Asexual parasites express many virulence factors and are the targets of anti-malarials such as chloroquine; hence an in-depth understanding of their transcriptional profiles will set the stage for future experiments addressing responses to immune or drug pressure.

SAGE was successfully applied to erythrocytic stage parasites (3D7 strain) of *P. falciparum* at baseline culturing conditions, and a SAGE library of approximately 8000 tags was generated. A majority of these corresponded to unique parasite genes, as demonstrated by BLAST analysis of a subset of tags. The SAGE data was validated by Northern and RT-PCR analysis of genes predicted to be highly expressed based on tag counts. BLAST analysis of highly abundant tags also provided insight into networks of major metabolic pathways that are utilized by the parasite under normal culture conditions. Finally, SAGE also revealed the presence of anti-sense transcription in the malarial parasite, a phenomena that has been previously missed by other methods of transcriptional analysis. This SAGE data was also validated by two independent methods, RT-PCR and Northern analysis; here anti-sense transcripts for genes expressed in asexual as well as sexual stage parasites were found. The biological role of anti-sense RNA in *Plasmodium* species is unclear; the phenomenon may be involved to translation control or may reflect

mechanisms of transcriptional initiation. In summary, SAGE in *Plasmodium* has revealed many facets of the basic functioning of the parasite in culture, and sets the stage for future comparisons of the transcriptional responses of *P. falciparum* to different stimuli.

Materials and Methods

Parasite culture and RNA extraction

3D7 strain parasites were maintained under standard culturing conditions (Trager and Jensen, 1976) with modifications as previously described (Munasinghe et al., 2000). Polyadenylated RNA was harvested from cultures at 8% parasitemia (1% rings, 5% trophozoites and 2% schizonts), and used in the SAGE procedure as previously described (Munasinghe et al., 2000).

Data analysis

SAGE tags from 3D7 asexual stages were analyzed using the SAGE software (Johns Hopkins University and Genzyme), which extracts 14bp tag counts from sequence files. In order to assign gene identity to each tag, the 3D7 experimental tag list was matched against a *P.falciparum* tag database. This database was created by extracting 14bp tags from *P.falciparum* sequence deposited in GenBank (as of July 13th 2000), as well as from a compiled database of recently deposited 3D7 genome sequence (obtained from the TIGR, Sanger and Stanford sequencing centers and compiled at the University of Pennsylvania as of July 26th 2000- kindly provided by Drs. Jessica Kissinger and David Roos). As the *P. falciparum* genome is not fully annotated, all potential SAGE tags from both sense and anti-sense strands were extracted (i.e., tags were extracted from each database in the "genomic mode" rather than the "cDNA mode").

The software output files are organized in such a way that matches to a single locus, matches to multiple loci, and no matches to database sequences can be readily determined. For genomic sequence that is annotated, it is possible to assign gene identification to each tag in the manner outlined above; however, most of the available *P. falciparum* genome sequence is not annotated. Therefore, the 187 most abundant tags (abundance level of greater than 4) were characterized by manual BLASTx analysis – see flow chart in Figure 1. Here, for tags derived from un-annotated reads, a 500-1000bp sequence surrounding the tag was translated in all 6 reading frames and compared to the entire NCBI protein database. 14bp tags that failed to match either database were analyzed using only the first 13bp of the tag sequence in the manner outlined above.

Reverse transcriptase PCR

RT-PCR was performed using the 3' RACE kit (Gibco-BRL) according to the manufacturer's protocols. First strand cDNA synthesis was primed with oligo $(dT)_{18}$, while PCR was performed using the gene specific primers described below.

Calmodulin (sense):	5' GTCCATCACCATCAATATCAGC 3'
Calmodulin (anti-sense):	5' CTAAGGAGTTAGGAACGGTCATG 3'
msp-3 (sense):	5' TTTTTGTGTTCTGGAACGCCTCCTCC 3'
msp-3 (anti-sense):	5' GCTTCCGAAGATGCTGAAAAAGCTGC 3'
pfg27/25 (sense):	5' TCTTGTCGTTCATGATACGCTTC 3'
pfg27/25 (anti-sense):	5' GTACAAAAGGATAGTGCCAAGCCC 3'
rap-1 (sense):	5' CTTTGAAGAAATCTCTGATTTCAGC 3'
rap-1 (anti-sense):	5' GCTTTAGAAGGTGTCTGTTCATATC 3'

PCR reactions were carried out according to the manufacturer's protocol (3' RACE kit, Gibco-BRL). Initial denaturation of the template occurred at 94° C for 3 minutes. Amplification was performed for 5 cycles at 94° C for 45 seconds, 50° C for 45 seconds and 72° C for 45 seconds, followed by 26 cycles of identical amplification where the annealing temperature was increased to 55° C. Finally, extension of partial PCR products was completed at 72° C for 6 minutes.

Strand-specific RT-PCR utilized 0.5 μg of mRNA per reaction and was performed with the express purpose of distinguishing sense mRNA from anti-sense mRNA (Yu et al., 1995). RT-PCR was performed using the 3' RACE kit (Gibco-BRL); however, first strand cDNA was primed with gene-specific primers that hybridize to either sense or anti-sense messages, rather than with an oligo$(dT)_{18}$ primer. The gene-specific primers are identical to the primers listed above. A tenth of the cDNA sample was PCR amplified, using the same set of gene-specific primers and amplification conditions described above.

PCR products were electrophoresed on 1.2% agarose gels. All resultant PCR products were cloned into the pCRII vector using the TA cloning kit (Invitrogen) and sequenced to confirm the identity of the amplified cDNA.

Northern blots

Northern analysis was performed according to standard protocols. Briefly 1μg of mRNA from 3D7 cultures was gel electrophoresed, blotted onto BA85 nitrocellulose membranes (Schleicher and Schuell) and probed with gene specific DNA probes. All probes (calmodulin, *msp-3, rap-1* and *pfg27/25*) were derived from the RT-PCR products described in the previous section. DNA probes were radiolabeled with α^{32}P dATP using

random hexanucleotides and the Klenow fragment of DNA polymerase. Blots were visualized by autoradiography.

For strand-specific northerns, 20 μg of total RNA were used per blot as described above. Probes for strand-specific Northern analysis were generated in the following manner. RT-PCR products of calmodulin and *msp-3* cDNAs (see previous sections) were cloned into the pCRII vector and then sub-cloned into pBluescript. The orientation of calmodulin and *msp-3* genes within pBluescript was determined by sequencing. The template for synthetic RNA corresponding to the sense strand was obtained by digestion of each pBluescript plasmid with XhoI, while the template for synthetic RNA corresponding to anti-sense RNA was obtained by digestion of the pBluescript plasmids with BamHI. Synthetic RNAs corresponding to the sense or anti-sense strand of either gene were obtained by *in vitro* transcription reactions (performed according to standard protocols). Strand specific RNA probes were also obtained under the same conditions in the presence of α^{32}P ATP.

Quantitative northern analysis was carried out for calmodulin and *msp-3* to determine whether the ratio of their transcripts was comparable to that determined by SAGE. Northern blots and gene specific DNA probes were prepared as described above. Known amounts of synthetic 300bp RNA fragments from each gene were run alongside the mRNA sample as markers for quantification. Blots were exposed to X-ray film (Kodak XO-MAT) such that the intensity of the signal was within the linear range of the film. Signal intensities for each of the transcripts in the mRNA sample were converted to molar amounts by reference to those of the synthetic RNAs. Signal intensities were measured by

scanning the X-ray film into Adobe Photoshop, and utilizing NIH Image software to quantify bands by pixel density.

P. gallinaceum gamete preparation and RNA isolation

P.gallinaceum parasites were propagated in White leghorn chickens by serial injection into wing veins. At parasitemias of 50-70%, blood was withdrawn by heart puncture. Gametogenesis was induced as described previously (Goonewardene et al., 1993), with the inclusion of xanthurenic acid (Sigma) at a final concentration of 50 μM in the exflagellation buffer. Gametes and zygotes were purified, also as described previously (Goonewardene et al., 1993), and $1X10^7$ cells were incubated at 25°C in Medium 199 (Gibco-BRL) and harvested for analysis at 0, 24, and 48 hours after isolation. Total RNA was isolated using Tri reagent (Molecular Research Center, Inc.) according to the manufacturer's protocol. Total RNA obtained from $1X10^7$ parasites was used for each RT-PCR reaction. Strand-specific RT-PCR was performed as described previously with the following primers:

pgs28 (sense): 5' CATCTAGCATAGTCAGCACAAGGTTTATTTG 3'

pgs28 (anti-sense): 5' CAAACGAAGATTATTTAGTCAAAC 3'

Results

3D7 SAGE tag library from asexual blood stage parasites

A total of 8335 SAGE tags were analyzed from the asexual blood stages of *P.falciparum*, 3D7 strain. A preliminary analysis showed that these 8335 tags corresponded to 4798 unique genes (Figure 2A). Of these, 1254 genes were present at an abundance of two or greater. The 537 tags expressed at abundance levels greater than or

equal to 20 tags (percentage frequency of 0.2) accounted for 6.4% of the total mRNA mass but only 0.3% (15) of the total number of unique genes. As expected, these abundance groups had the highest percentage of matches to GenBank entries (Figure 2B), implying that many highly expressed messages have been readily cloned and studied. The lower abundance tags (frequency of less than 20 tags) accounted for 93.6% of the total mRNA mass, and represented a vast majority of the unique genes expressed in the parasite. Moreover, these tags gave many fewer matches to GenBank; hence SAGE in *Plasmodium falciparum* will aid in the discovery of novel malarial genes.

BLAST analysis of SAGE tags

To assess whether 14bp tags could uniquely identify genes in the highly A-T rich *Plasmodium* genome, these SAGE tags were searched against 3D7 genome sequence. We decided that for an accurate estimate of the "tag to gene" mapping in *Plasmodium*, all available sequence data, both cDNA and genomic, would provide the most complete picture. Sequencing of the *P. falciparum* genome is close to completion; however, much of the newly available *P. falciparum* sequence data has yet to be annotated. Therefore, the 187 most abundant SAGE tags were analyzed in a more rigorous manner by BLASTx analysis. A schematic of the BLAST analysis is shown in Figure 1. This analysis revealed that a majority of the SAGE tags (88%) corresponded to *P. falciparum* genome sequence. Most of the tags that match to single loci (70%) lie within known genes; hence SAGE tags can be used to uniquely identify genes in *Plasmodium*. The other 30% of tags that match single sites correspond to unknown genes and hypothetical open reading frames. Thus SAGE data reveals not only predicted ORFs that are expressed but also previously

uncharacterized ORFs; hence SAGE in *Plasmodium* has the capacity to assist in annotation of the genome.

Approximately 10% of the 187 most abundant SAGE tags did not match parasite sequence. We expect this number to decrease as the genome project nears completion. The percentage of SAGE tags that gave multiple matches within the *P. falciparum* genome was also calculated and found to be 18%. This number is about four-fold higher than that obtained by Velculescu *et al* (1995) in a SAGE library of human pancreatic tissue. In the present study, the 35 tags that matched more than one loci were further investigated; of these tags, 21 (60%) matched 2 or 3 genes while 14 (40%) matched greater than 3 genes. The latter set of tag sequences was of lower complexity in general. Northern Blot analysis should help resolve whether tags that match multiple genes indeed represent multiple transcripts.

Abundant transcripts expressed in *P. falciparum* grown in culture

The BLAST analysis described earlier enabled us to assign genes to highly abundant SAGE tags; examples of these are listed in Table 1. This analysis provided a snapshot of the major transcripts expressed by the parasite. A complete picture of metabolic pathways utilized by *P. falciparum* growing in culture will incorporate protein expression and stability; nevertheless, BLAST analysis of abundant SAGE tags provides the first global description of genes and hence, metabolic pathways that might be transcriptionally regulated. The most abundant transcripts were grouped into functional categories to reveal the transcriptional profile of 3D7 parasites grown in culture (Figure 3). Many tags represented housekeeping functions carried out by all prokaryotic and eukaryotic cells (transcription, translation, chaperones, cytoskeleton, etc.) while some functional classes

were highly specific for the unique life cycle of *Plasmodium* (membrane associated proteins involved in invasion, DOXP pathway).

Interestingly, many of the most abundant messages (5.3%) appear to be transcribed from the 6kb mitochondrial genome and another 2.1% (thioredoxin, vacuolar ATPase subunit B, ATPase transporter, ubiquinol cytochrome-c reductase like protein) are required for oxidative metabolism. Therefore a significant proportion of abundant transcripts encode proteins that are dedicated towards redox processes.

Stage-specific transcripts are highly represented in the list of abundant messages, reflecting the different developmental stages present in the culture. For example, mRNAs encoding cell surface proteins involved in merozoite invasion (Cowman et al., 2000) comprise 8% of the most abundant transcripts. These include merozoite surface proteins 3 and 4 (MSP-3 and -4), rhoptry associated protein-1 (RAP-1) and merozoite capping protein. Tags corresponding to serine repeat antigen, a soluble protein that is associated with the parasitophorous vacuole were found at high abundance (0.32%). Also, present at high abundance (0.25%) is a tag representing the gametocyte surface antigen Pfg27/25, shown to be essential for gametogenesis (Lobo et al., 1999).

Abundant SAGE tags represented major metabolic pathways of the malarial parasite. As asexual blood stages of *Plasmodium* do not store energy reserves in the form of glycogen or lipids, glucose taken up from plasma is the primary source of energy (Sherman, 1991). Therefore, glucose metabolism is a prominent aspect of intracellular growth and not unexpectedly, proteins required for glucose metabolism were represented among the abundant tags (aldolase, PEP carboxykinase, and triosephosphate isomerase).

Although lipids are not utilized as a major source of energy by *P. falciparum*, there is a significant increase in levels of phospholipids, diacylglycerol and triacylglycerol within the red blood cell upon merozoite invasion (Vial and Ancelin, 1998). This increase in the total lipid content is associated with a biosynthetic requirement for lipids during formation of the membranes surrounding the parasite (the parasitophorous vacuolar membrane and the tubovesicular membrane). N-myristoyl transferase, an enzyme that plays a role in the formation of lipoproteins, was found among the 187 most abundant tags; however, tags representing proteins involved in lipid biosynthesis were not present.

Intra-erythrocytic *P. falciparum* parasites are capable of *de novo* synthesis of pyrimidines from precursor molecules (Walsh and Sherman, 1968), with a requirement for para-aminobenzoic acid (pABA) and folate cofactors. Unlike their hosts, malarial parasites do not use exogenous folate cofactors, but instead, synthesize these *de novo* (Scheibel and Sherman, 1988). SAGE data revealed tags corresponding to ribonucleotide reductase, an enzyme of the pyrimidine biosynthetic pathway and dihydrofolate synthase, an enzyme of the folate pathway. Polyamine biosynthetic enzymes were also represented among the SAGE tags (ornithine decarboxylase and ornithine aminotransferase).

The unique intracellular niche of malarial parasites results in the expression of many parasite-specific metabolic pathways. For example, growth of the asexual parasites within red blood cells is accompanied by degradation of hemoglobin and the subsequent detoxification of heme by-products (Foley and Tilley, 1998; Krogstad and De, 1998; Rosenthal and Meshnick, 1998). Tags representing proteins implicated in the detoxification of heme (histidine-rich proteins I and II, glutathione reductase) were found at high abundance in the SAGE library. Surprisingly, the plasmepsin and falcipain

proteases, that play a role in hemoglobin degradation, were not found in the list of highly expressed genes. This may be related to the protein stability of these factors or may be due to the fact that their transcription occurs at an earlier stage in the parasite life cycle than the trophozoite stage, which was the predominant stage in the study population.

Finally, SAGE data revealed the expression of mRNA encoding deoxy-D-xylulose 5-phosphate synthase (DOXP synthase) at high levels (0.09%). The DOXP pathway was recently identified as a parasite-specific metabolic pathway important for isoprenoid biosynthesis (Jomaa et al., 1999). As this pathway is localized in the apicoplast, a plant-derived organelle of *Plasmodium*, DOXP metabolism provides a novel target for anti-malarial drug development.

Validation of SAGE data

In order to confirm the expression data in asexual stage parasites as determined by SAGE, RT-PCR and Northern analysis of several genes with highly abundant SAGE tag counts (calmodulin, *msp-1*, *rap-1*, and *pfg27/25*- see Figure 4) were performed. Pfg27/25 represents a gametocyte-specific antigen, while the other three are predicted to be expressed in asexual stages. As the SAGE library was derived from a culture that contained no detectable gametocytes, *pfg27/25* was specifically chosen for RT-PCR and Northern analysis. RT-PCR products for all four genes were generated from asexual stage mRNA (Figure 4A). These were cloned, sequenced and found to correspond to the expected gene. Transcripts at the predicted length for all four genes were also detected by Northern blotting (data not shown, see Figure 4B)

For a more quantitative estimate of gene expression, quantitative Northern analysis of two highly expressed genes (*msp-3* and calmodulin) was performed (data not shown,

Figure 4B). Here the molar ratio of *msp-3* to calmodulin was approximately 3:1, which is similar to the ratio of their SAGE tag counts (Figure 4B). Hence, SAGE tag data appears to correlate well with levels of mRNA within the cells.

Anti-sense transcripts

A surprising observation of SAGE in *P. falciparum* was the large proportion of tags corresponding to anti-sense transcripts. Unlike microarrays, SAGE is able to detect anti-sense transcription since the orientation of the SAGE tag on the mRNA can be readily determined. A SAGE tag consist of the 4bp recognition sequence (CATG) of the restriction enzyme, NlaIII (this enzyme defines the position of each tag in a mRNA transcript) and 10bp of adjacent sequence in the direction of the 3' poly A tail of the RNA molecule. Among 45 annotated genes whose 5' and 3' ends are clearly denoted, 17% of the tags consisted of a CATG and the 3' adjacent 10bp, in the direction of the 5' end of the transcript, on the non-coding strand of cDNA. This result was unexpected; hence, we wanted independent confirmation of the SAGE data. This was accomplished by strand specific RT-PCR analysis of asexual as well as sexual blood stages, and strand specific Northern analysis in erythrocytic stage parasites.

We confirmed the presence of anti-sense transcripts from erythrocytic stages by strand-specific RT-PCR analysis of three genes, calmodulin, *rap-1* and *msp-3*; and subsequent sequencing of the RT-PCR products to establish gene identity. Based on SAGE data, all three transcripts were expected to be present in both the sense and anti-sense orientations, a prediction that was confirmed by RT-PCR (Figure 5A, lanes 5-16) and sequence analysis. Although a PCR product was detected for *pfg27/25* anti-sense RNA (Figure 5A, lane1), the sequence of the PCR product did not correspond to

pfg27/25, consistent with the absence of an anti-sense SAGE tag for this gene. Importantly, control experiments that excluded reverse transcriptase (lanes 2,4,6,8,10,12,14,16) indicated a lack of contaminating genomic DNA, showing that the PCR products obtained during strand-specific RT-PCR were indeed derived from RNA. These data validate the anti-sense transcripts predicted by SAGE.

The presence of anti-sense transcripts was also confirmed by strand-specific Northern analysis for calmodulin and *msp-3*. Figure 5B shows that strand-specific probes can specifically detect synthetic anti-sense RNA (lanes 1 and 2 for calmodulin; lanes 7 and 8 for *msp-3*) or synthetic sense RNA (lanes 4 and 5 for calmodulin; lanes 10 and 11 for *msp-3)*. Using these strand-specific probes, total RNA isolated from asexual stage parasites was shown to contain both anti- sense and sense transcripts for both calmodulin (lanes 3 and 6) and *msp-3* (lanes 9 and 12). Therefore, as confirmed by two independent techniques, the presence of anti-sense tags in the SAGE library reflects anti-sense transcription in asexual stages of the malarial parasite.

We wondered whether genes expressed in other stages of the *Plasmodium* life cycle also exhibited anti-sense transcription. To address this, the sexual stages (zygotes and ookinetes) of the chicken malarial parasite, *P. gallinaceum*, were tested for the presence of anti-sense RNAs. Pgs28 is a major surface antigen of *P. gallinaceum* sexual stages (Duffy et al., 1993) and transcription of the *pgs28* gene has been studied previously (29). Strand-specific RT-PCR of total RNA from zygotes (0 hours) and mature ookinetes (48 hours) showed that the *pgs28* gene expressed both sense and anti-sense transcripts (Figure 6) at different stages of *in vitro* development.

Discussion

This report demonstrates the application of SAGE in *P. falciparum*. Despite the low complexity of the genome, SAGE tags as short as 14bp can uniquely identify a majority of genes in *P. falciparum*. This observation has been exploited to study transcription in the asexual stages of the parasite, resulting in new insights into the biology of the pathogen. First, we provide a description of the transcriptional profile of the 3D7 strain of *P. falciparum* that builds upon the extensive data generated by the Malaria Genome Project. Second, the major metabolic pathways present in blood stage parasites are delineated; modulation of these pathways in response to stimuli like drug- and immune-pressure can now be studied. And finally, this report shows that *Plasmodium* parasites express anti-sense RNAs at multiple stages during the developmental cycle, a finding that has implications for transcriptional regulation of *Plasmodium* gene expression.

Analysis of SAGE tags

Of the tags that matched to single loci, 70% matched to known genes while 30% matched to unknown genes or hypothetical proteins. This distribution is in stark contrast to genome sequencing data where 60% of the putative ORFs were of unknown function while 40% were genes encoding proteins of known functions (Gardner et al., 1998). This discrepancy could be explained by the fact that the asexual blood stages are more amenable to cultivation and experimental manipulation in the laboratory than other stages; hence, many of the transcripts expressed in these stages have been previously studied and are of known functions. It is also likely that a majority of the transcripts expressed during laboratory culture of asexual blood stages encode proteins that serve housekeeping functions conserved within organisms widely separated on the phylogenetic tree. The

genes of unknown function identified by the Malaria Genome Project may turn out to be of importance in host-parasite interactions and disease; however, under culturing conditions only relatively few may be expressed at high levels. Moreover, as SAGE data reveals genes that are actually expressed in asexual stage parasites, identification of tags that correspond to unknown genes and hypothetical proteins will be of tremendous use in annotation of the *P. falciparum* genome.

Some tags (10%) did not match to the *Plasmodium* databases. As the *P. falciparum* genome is 80-90% complete, these tags should prove to be informative as the genome project proceeds to completion. Alternatively, tags that do not match genome sequence may turn out to span splice junctions. These questions should be resolved, as more genome sequence becomes available. Nevertheless, SAGE in *P. falciparum* is comparable to other studies where tags with no matches to the genome were as high as 20% (Matsumura et al., 1999) and 23% (Yamashita et al., 2000) of the total tags.

Finally, of the 8335 tags, 18% gave multiple matches to *Plasmodium* databases, a number that is four-fold higher than that obtained from human pancreatic SAGE libraries, where ~5% of tags gave multiple matches (Velculescu et al., 1995). However, pancreatic SAGE tags were only searched against RNA sequence databases, in contrast to our more extensive analysis that surveyed all available *Plasmodium* genome sequence. Hence, the higher percentage of multiple matches to the genome may reflect the method of analysis rather than any limitation of the technique when applied to the A-T rich genome of *Plasmodium*. Alternatively, the higher percentage of tags giving multiple matches may be a consequence of the lower complexity of the *Plasmodium* genome. Ambiguous tags of interest can be investigated further on an individual basis by Northern analysis.

Metabolic pathways defined by SAGE

Other reports on SAGE have revealed metabolic profiles that are highly specific to the organism or tissue under study. For example, SAGE of mouse kidney revealed a preponderance of ion channels and mitochondrial enzymes, consistent with the role of the kidney in filtration and solute transport and the high energy requirement for the same (El-Meanawy et al., 2000). Transcriptional profiling of the 100 most abundant SAGE tags derived from seedlings of the rice plant, *Oryza sativa* L., demonstrated a prevalence of prolamin, a storage protein expressed in seeds (Matsumura et al., 1999). Not unexpectedly, other highly abundant transcripts included those encoding water channels and respiratory metabolism enzymes.

SAGE data from *P. falciparum* sheds light on the transcriptional profile of blood stage parasites and hence reveals the classes of proteins and metabolic pathways that are utilized during asexual growth. For example, membrane-associated proteins form the most abundant category of expressed proteins. This is not surprising in light of the fact that the parasite is separated from its extracellular environment by three separate membranes: the host red blood cell membrane, the parasitophorous vacuole membrane, and the parasite plasma membrane. Many of these highly expressed proteins are stage specific and have been previously shown to be important in invasion of the red blood cell (merozoite surface proteins-3 and -4); others are transporters that may import nutrients into the parasite cell (importin β-subunit). Hence, the unique niche of the malarial parasite within the red blood cell requires the high expression of specific surface proteins.

A significant proportion (7.4%) of the most abundant tags were derived either from transcripts encoded on the 6kb mitochondrial genome or from nuclear encoded

transcripts involved in redox metabolism. High levels of RNA synthesis from the 6kb element may reflect the fact that this episomally replicating molecule is present at approximately 20 copies per cell (Preiser et al., 1996). However, the high abundance of nuclear encoded transcripts also involved in redox metabolism (thioredoxin, ubiquinol cytochrome c-reductase like protein) indicates that a large proportion of the cells metabolic activities involve the maintenance of intracellular oxidative homeostasis. Moreover, SAGE data show that transcripts encoding the molecular chaperones Hsp-60 and -70, which may be involved in import of nuclear encoded proteins into the mitochondria (Das et al., 1997), are also expressed at high levels. Hence, mitochondrial functions are most highly represented in the abundant classes of SAGE tags, likely reflecting the micro-aerophilic lifestyle of the parasite within the red blood cell. The robust expression of genes involved in mitochondrial physiology may explain why mitochondrial pathways have been excellent targets for anti-malarial drugs.

The major transcriptional pathways in the parasite as revealed by SAGE will help to identify potential drug targets and lead compounds. For example, atovaquone inhibits erythrocytic growth by targeting the mitochondrial cytochrome bc_1 complex (Fry and Pudney, 1992). Further evidence that other highly expressed metabolic pathways could also serve as drug targets is found in the following studies: the anti-malarial drug fosmidomycin has been shown to target DOXP metabolism (Jomaa et al., 1999); the ornithine decarboxylase inhibitor, difluoro-methylornithine, inhibits erythrocytic growth of *P. falciparum* in culture (Assaraf et al., 1984); and folate antagonists like pyrimethamine and cycloguanil target dihydrofolate reductase (Ferone et al., 1969). Other major

transcriptional patterns uncovered by SAGE in the parasite (proteasome, chaperones, unknown ORF, etc.) may provide new targets for anti-malarial drug development.

SAGE reveals novel transcriptional phenomena in *P. falciparum*

Most techniques for global analysis of gene expression are unable to distinguish sense and anti-sense transcripts. Due to the directional nature of SAGE tags (3' most NlaIII site of each transcript (4bp) and 10bp downstream on the coding strand) we were able to identify numerous anti-sense transcripts in the transcriptional repertoire of *P. falciparum* asexual stage parasites. Strand specific RT-PCR and Northern analysis confirmed this observation for three of the genes (*msp-3, rap-1* and calmodulin) predicted to transcribe anti-sense messages. The fact that anti-sense transcription can be detected in *Plasmodium* by three independent methods suggests that this a *bona fide* biological phenomenon and not an artifact of the SAGE procedure. It is of interest to note that anti-sense transcripts in genes that contain introns are larger than their corresponding sense RNAs suggesting that sequences complementary to introns are present in the former.

Our data also demonstrates that anti-sense transcripts are expressed in other stages of *Plasmodium* development. The *pgs28* gene that encodes a major surface antigen of *P. gallinaceum* sexual stages has been studied extensively (Duffy et al., 1993). Transcription of *pgs28* is restricted to the zygotes and ookinetes. Strand-specific RT-PCR shows that *pgs28* expresses both sense and anti-sense transcripts in both stages. Hence, the presence of anti-sense transcripts may be a widespread phenomenon in multiple stages of *Plasmodium* development and should be tested further. For example, a family of genes (*var*) encodes variable surface proteins involved in host-parasite interactions; *var* genes are transcribed during erythrocytic growth resulting in the expression of the PfEMP-1

protein (Wahlgren et al., 1999). Several *var* genes are transcribed in the ring stages while a single *var* gene is transcribed in trophozoites (Chen et al., 1998; Scherf et al., 1998). It would be interesting to test whether any of the ring stage *var* transcripts are anti-sense.

Anti-sense transcripts may reflect mechanisms of transcriptional initiation in a parasite with a highly A-T rich genome (86% A-T in non-coding regions and 76% in coding sequence) (Bowman et al., 1999). Numerous studies have shown that transcription in *P. falciparum* is initiated from the A-T rich 5' upstream region of genes resulting in sense transcripts (Dechering et al., 1999; Horrocks et al., 1998; Horrocks and Lanzer, 1999). Sense and anti-sense message derived from genes that do not contain introns were approximately the same size, suggesting that transcription may also initiate from the 3'downstream intergenic region of genes. The presence of anti-sense transcripts for 17% of annotated genes implies novel mechanisms of transcriptional initiation and termination, including potential roles in post-transcriptional control of protein expression.

In conclusion, we have shown that SAGE can be readily adapted for the study of global transcription in *Plasmodium falciparum*. SAGE of 3D7 asexual parasites sheds light on the prominent metabolic pathways utilized in these stages. Since blood stages are the targets of both anti-malarial drugs and the host immune system, this comprehensive transcriptional profile generated by SAGE will form the basis for future comparisons of gene expression under drug or immune pressure. Finally, the unique nature of SAGE reveals novel phenomena, that of anti-sense transcription which has previously been missed.

Acknowledgements

We would like to thank Drs. J. Kissinger and D. Roos (University of Pennsylvania; droos@sas.upenn.edu) for providing access to assembled *P. falciparum* genomic sequences.

We acknowledge the invaluable data provided by the Malaria Genome Consortium: Sequence data for *P. falciparum* chromosome (1,3,4,5,6,7,8,9,13) was obtained from The Sanger Centre website at http://www.sanger.ac.uk/Projects/P_falciparum/. Sequencing of *P. falciparum* chromosome (1,3,4,5,6,7,8,9,13) was accomplished as part of the Malaria Genome Project with support by The Wellcome Trust. Sequence data for *P. falciparum* chromosome 12 was obtained from the Stanford DNA Sequencing and Technology Center website at http://www-sequence.stanford.edu/group/malaria. Sequencing of *P. falciparum* chromosome 12 was accomplished as part of the Malaria Genome Project with support by the Burroughs Wellcome Fund. Preliminary sequence data for *P. falciparum* chromosome (2,10,11,14) was obtained from The Institute for Genomic Research website (www.tigr.org). Sequencing of chromosome (2,10,11,14) was part of the International Malaria Genome Sequencing Project and was supported by awards from the Burroughs Wellcome Fund and the U.S. Department of Defense. The Chromosome 2 Sequencing Project was a collaborative effort by The Institute of Genomic Research (TIGR) and the Naval Medical Research Center (NMRC).

We also thank Dr. Connie Chow for insightful comments about this manuscript. This work was supported by the Burroughs Wellcome Fund and the Department of Defense.

References

Assaraf, Y.G., Golenser, J., Spira, D.T. and Bachrach, U. (1984). Polyamine levels and the activity of their biosynthetic enzymes in human erythrocytes infected with the malarial parasite, Plasmodium falciparum. Biochem J. *222*, 815-819.

Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C.M., Craig, A., Davies, R.M., Devlin, K., Feltwell, T., Gentles, S., Gwilliam, R., Hamlin, N., Harris, D., Holroyd, S., Hornsby, T., Horrocks, P., Jagels, K., Jassal, B., Kyes, S., McLean, J., Moule, S., Mungall, K., Murphy, L., Barrell, B.G. and et al. (1999). The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum [see comments]. Nature. *400*, 532-538.

Butler, D. (1997). Funding assured for international malaria sequencing project [news]. Nature. *388*, 701.

Chen, Q., Fernandez, V., Sundstrom, A., Schlichtherle, M., Datta, S., Hagblom, P. and Wahlgren, M. (1998). Developmental selection of var gene expression in Plasmodium falciparum. Nature. *394*, 392-395.

Cowman, A.F., Baldi, D.L., Healer, J., Mills, K.E., O'Donnell, R.A., Reed, M.B., Triglia, T., Wickham, M.E. and Crabb, B.S. (2000). Functional analysis of proteins involved in Plasmodium falciparum merozoite invasion of red blood cells. FEBS Lett. *476*, 84-88.

Craig, A.G., Waters, A.P. and Ridley, R.G. (1999). Malaria genome project task force: a post-genomic agenda for functional analysis [news]. Parasitol Today. *15*, 211-214.

Das, A., Syin, C., Fujioka, H., Zheng, H., Goldman, N., Aikawa, M. and Kumar, N. (1997). Molecular characterization and ultrastructural localization of Plasmodium falciparum Hsp 60. Mol Biochem Parasitol. *88*, 95-104.

Dechering, K.J., Kaan, A.M., Mbacham, W., Wirth, D.F., Eling, W., Konings, R.N. and Stunnenberg, H.G. (1999). Isolation and functional characterization of two distinct sexual-stage- specific promoters of the human malaria parasite Plasmodium falciparum. Mol Cell Biol. *19*, 967-978.

Duffy, P.E., Pimenta, P. and Kaslow, D.C. (1993). Pgs28 belongs to a family of epidermal growth factor-like antigens that are targets of malaria transmission-blocking antibodies. J Exp Med. *177*, 505-510.

El-Meanawy, M.A., Schelling, J.R., Pozuelo, F., Churpek, M.M., Ficker, E.K., Iyengar, S. and Sedor, J.R. (2000). Use of serial analysis of gene expression to generate kidney expression libraries [In Process Citation]. Am J Physiol Renal Physiol. *279*, F383-392.

Ferone, R., Burchall, J.J. and Hitchings, G.H. (1969). Plasmodium berghei dihydrofolate reductase. Isolation, properties, and inhibition by antifolates. Mol Pharmacol. *5*, 49-59.

Foley, M. and Tilley, L. (1998). Quinoline antimalarials: mechanisms of action and resistance and prospects for new agents. Pharmacol Ther. *79*, 55-87.

Fry, M. and Pudney, M. (1992). Site of action of the antimalarial hydroxynaphthoquinone, 2-[trans-4- (4'-chlorophenyl) cyclohexyl]-3-hydroxy-1,4-naphthoquinone (566C80). Biochem Pharmacol. *43*, 1545-1553.

Gardner, M.J., Tettelin, H., Carucci, D.J., Cummings, L.M., Aravind, L., Koonin, E.V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai, Z., Schwartz, D.C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G.G., Clayton, R., White, O., Smith, H.O., Fraser, C.M., Hoffman, S.L. and et al. (1998). Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum [published erratum appears in Science 1998 Dec 4;282(5395):1827]. Science. *282*, 1126-1132.

Goonewardene, R., Daily, J., Kaslow, D., Sullivan, T.J., Duffy, P., Carter, R., Mendis, K. and Wirth, D. (1993). Transfection of the malaria parasite and expression of firefly luciferase. Proc Natl Acad Sci U S A. *90*, 5234-5236.

Hayward, R.E., Derisi, J.L., Alfadhli, S., Kaslow, D.C., Brown, P.O. and Rathod, P.K. (2000). Shotgun DNA microarrays and stage-specific gene expression in plasmodium falciparum malaria [In Process Citation]. Mol Microbiol. *35*, 6-14.

Horrocks, P., Dechering, K. and Lanzer, M. (1998). Control of gene expression in Plasmodium falciparum. Mol Biochem Parasitol. *95*, 171-181.

Horrocks, P. and Lanzer, M. (1999). Mutational analysis identifies a five base pair cis-acting sequence essential for GBP130 promoter activity in Plasmodium falciparum. Mol Biochem Parasitol. *99*, 77-87.

Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K., Soldati, D. and Beck, E. (1999). Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science. *285*, 1573-1576.

Krogstad, D. and De, D. (1998) Chloroquine: modes of action and resistance and the activity of chloroquine analogs. In Sherman, I.W. (ed.) *Malaria: parasite biology, pathogenesis and protection*. ASM Press, Washington, DC, pp. 331-339.

Lobo, C.A., Fujioka, H., Aikawa, M. and Kumar, N. (1999). Disruption of the Pfg27 locus by homologous recombination leads to loss of the sexual phenotype in P. falciparum. Mol Cell. *3*, 793-798.

Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H. and Beaudry, G.A. (1997). SAGE transcript profiles for p53-dependent growth regulation. Oncogene. *15*, 1079-1085.

Matsumura, H., Nirasawa, S. and Terauchi, R. (1999). Technical Advance: Transcript profiling in rice (Oryza sativa L.) seedlings using serial analysis of gene expression (SAGE). Plant J. *20*, 719-726.

Munasinghe, A., Patankar, S., Cook, B.P., Madden, S.L., Martin, R.K., Kyle, D.E., Shoaibi, A., Cummings, L.M. and Wirth, D.F. (2000). Serial Analysis of Gene Expression (SAGE) in Plasmodium falciparum: application of the technique to A-T rich genomes. Mol Biochem Parasitol. *In press*,

O'Brien, C. (1997). Malaria genome project gets a funding boost [news]. Mol Med Today. *3*, 3.

Polyak, K., Xia, Y., Zweier, J.L., Kinzler, K.W. and Vogelstein, B. (1997). A model for p53-induced apoptosis [see comments]. Nature. *389*, 300-305.

Preiser, P.R., Wilson, R.J., Moore, P.W., McCready, S., Hajibagheri, M.A., Blight, K.J., Strath, M. and Williamson, D.H. (1996). Recombination associated with replication of malarial mitochondrial DNA. Embo J. *15*, 684-693.

Rosenthal, P. and Meshnick, S. (1998) Hemoglobin processing and the metabolism of amino acids, heme and iron. In Sherman, I.W. (ed.) *Malaria: parasite biology, pathogenesis and protection.* ASM Press, Washington DC, pp. 145-158.

Scheibel, L. and Sherman, I.W. (1988) Plasmodial metabolism and related organellar function during various stages of the life cycle: proteins, lipids, nucleic acids and vitamins. In Wernsdorfer, W. and McGregor, I. (eds.), *Malaria, Principles and Practice of Malariology.* Churchill Livingstone, Ltd., Edinburgh, United Kingdom., pp. 219-252.

Scherf, A., Hernandez-Rivas, R., Buffet, P., Bottius, E., Benatar, C., Pouvelle, B., Gysin, J. and Lanzer, M. (1998). Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. Embo J. *17*, 5418-5426.

Sherman, I.W. (1991) The biochemistry of malaria: an overview. In Coombs, G. and North, M. (eds.), *Biochemical Protozoology*. Taylor and Francis, London, United Kingdom, pp. 6-34.

Trager, W. and Jensen, J.B. (1976). Human malaria parasites in continuous culture. Science. *193*, 673-675.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995). Serial analysis of gene expression [see comments]. Science. *270*, 484-487.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr., Hieter, P., Vogelstein, B. and Kinzler, K.W. (1997a). Characterization of the yeast transcriptome. Cell. *88*, 243-251.

Vial, H.J. and Ancelin, M.L. (1998) Malarial Lipids. In Sherman, I.W. (ed.) *Malaria: Parasite Biology, Pathogenesis and Protection*. ASM Press, Washington D.C., pp. 159-175.

Virlon, B., Cheval, L., Buhler, J.M., Billon, E., Doucet, A. and Elalouf, J.M. (1999). Serial microanalysis of renal transcriptomes. Proc Natl Acad Sci U S A. *96*, 15286-15291.

Wahlgren, M., Fernandez, V., Chen, Q., Svard, S. and Hagblom, P. (1999). Waves of malarial var-iations. Cell. *96*, 603-606.

Walsh, C.J. and Sherman, I.W. (1968). Purine and pyrimidine synthesis by the avian malaria parasite, *Plasmodium lophurae*. Journal of Protozoology. *15*, 763-770.

WHO. (1997) The World Health Report. Conquering, Suffering, Enriching Humanity. WHO publishers, Geneva.

Yamashita, T., Hashimoto, S., Kaneko, S., Nagai, S., Toyoda, N., Suzuki, T., Kobayashi, K. and Matsushima, K. (2000). Comprehensive gene expression profile of a normal human liver. Biochem Biophys Res Commun. *269*, 110-116.

Yu, D.C., Wang, A.L., Wu, C.H. and Wang, C.C. (1995). Virus-mediated expression of firefly luciferase in the parasitic protozoan Giardia lamblia. Mol Cell Biol. *15*, 4867-4872.

Figure legends

Figure 1: BLASTx analysis of highly abundant SAGE tags. SAGE tags from the 3D7 library were analyzed using the SAGE software (as described in the methods section). SAGE tags at an abundance level of 5 or greater were further analyzed as depicted in the flow chart. The percentage of single matches, no matches and multiple matches to the databases are indicated. Numbers in brackets correspond to the proportion of tags counts in each group. Single matches were divided into tags that matched Annotated and Un-annotated sequence reads. Tags in the latter group were characterized by BLASTx. 14bp tags that failed to match either database were analyzed using only the first 13bp of the tag sequence.

Fig 2: SAGE 3D7 library analysis. *A. Cumulative total gene representation within the 3D7 SAGE library.* Ascertained tags (from the 3D7 library) at increasing increments were plotted against the number of unique genes from which the tag subsets were derived. The solid line corresponds to all ascertained tags. The dotted line corresponds to the number of unique genes represented by tags at an abundance level of 2 or greater (some tags at an abundance level of 1 may be derived from sequencing errors). *B. SAGE abundance classes.* 8335 SAGE tags are divided into abundance classes. The number of unique tags matching an entry in the *P.falciparum* NCBI database is listed per abundance class (last column), and the percentage of hits within the abundance class is given in brackets.

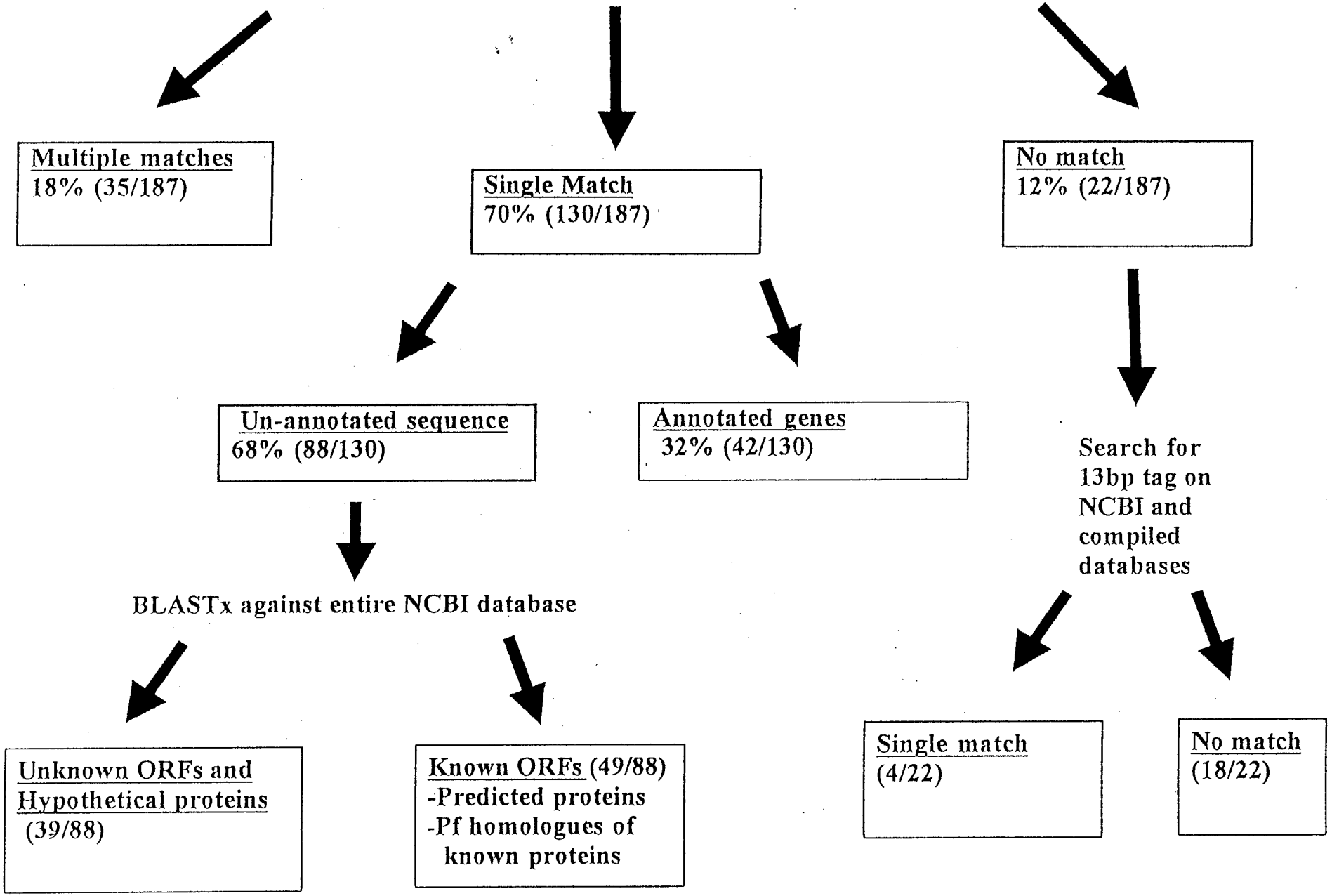
Fig 3 : Categories of highly expressed genes in the 3D7 control population. Highly abundant tags (percentage frequency of greater than 0.05) were examined for their matches in *P.falciparum* databases (see analysis described in figure 1) and categorized by putative functions. The number of genes in each category is depicted.

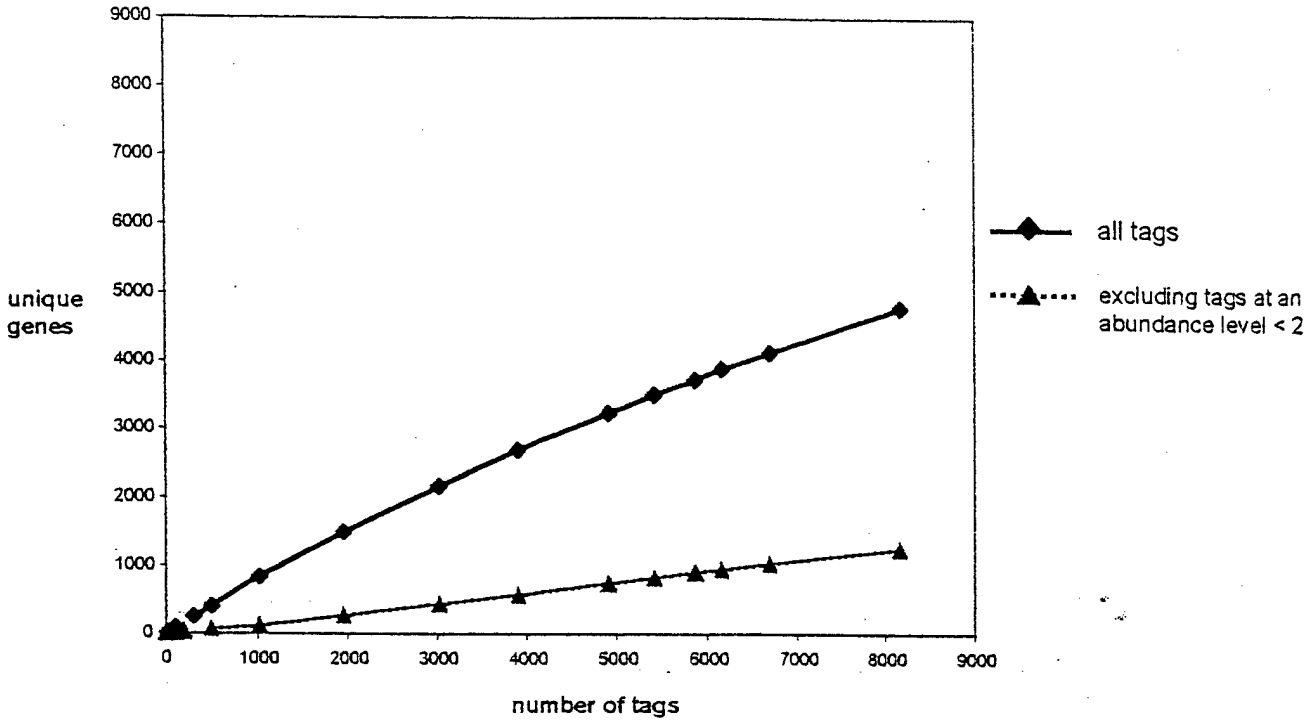
Figure 4 : Validation of SAGE data by RT-PCR, and Northern analysis *A. RT-PCR of genes represented in the 3D7 SAGE library.* RT-PCR products generated by specific primers for *pfg27/25* (lane 1 and 2), *rap-1* (lane 3 and 4), *msp-3* (lane 5 and 6) and calmodulin(lane 7 and 8) are shown. RT minus samples were also PCR amplified as negative controls (lanes 2,4,6,and 8). pBR322 MspI digest (lane M). *B. Summary of expression data.* For Northern analysis 1μg of mRNA from 3D7 cultures was gel electrophoresed, blotted onto a nylon membrane and probed with gene specific ^{32}P-labeled DNA probes. + indicates that a specific signal corresponding to each transcript was detected in the northern analysis. Tag counts from the 3D7 library (8335 tags) for all four genes are listed. Quantitative northern analysis was carried out for calmodulin and *msp-3*. Northern blots and gene specific probes were prepared as described above. Known amounts of *in-vitro* synthesized RNA fragments were run alongside the mRNA sample as markers for quantification. Signal intensities for each of the transcripts in the mRNA sample were converted to molar amounts by reference to those of the synthetic RNAs.

Fig 5: Validation of anti-sense SAGE data by stand specific RT-PCR and strand specific northern analysis in asexual stage parasites. *A. Strand specific RT-PCR analysis.* First strand cDNA from asexual stages was synthesized with a primer that specifically hybridizes to either anti-sense message (A) for *pfg27/25* (lane 1 and 2), *rap-1* (lane 5 and 6), *msp-3* (lane 9 and 10) and calmodulin (lane 13 and 14); or a primer that binds to sense message (S) for *pfg27/25* (lane 3 and 4), *rap-1* (lane 7 and 8), *msp-3* (lane 11 and 12) and calmodulin (lane 15 and 16). RT-PCR was performed on these cDNA

samples using the same primer pair for *pfg27/25* (lane 1-4), *rap-1* (lane 5-8), *msp-3* (9-12) and calmodulin(lane 13-16). Products in lanes 1,3,5,7,9,11,13,and 15 were cloned to confirm gene identity. "+" indicates that the gel product corresponded to the specific gene under investigation. Only the product in lane 1 was generated by non-specific PCR amplification, indicated by "-". RT minus samples were also PCR amplified as negative controls (lanes 2,4,6, 8,10, 12,14, and 16). pBR322 MspI digest (lanes M). Genes for which anti-sense transcripts were found in the SAGE library are indicated by "yes" or "no" (last row). *B. Strand specific northern analysis.* Synthetic sense (S) RNA fragments (lane 1 and 4 for calmodulin; lane 7 and 10 for *msp-3*), or synthetic anti-sense (A) RNA fragments (lane 2 and 5 for calmodulin; lane 8 and 11 for *msp-3*), or 20μg of total RNA from 3D7 cultures (lanes 3,6,9 and 12) were gel electrophoresed. Blots were transferred onto nitrocellulose membranes and probed with ^{32}P labeled sense (S) RNA probes (calmodulin: lane 1-3; *msp-3*: lane 7-9) or anti-sense (A) RNA probes (calmodulin: lane 4-6; *msp-3*: lane 10-12).

Fig 6: Strand specific RT-PCR analysis of *pgs28* in sexual stage parasites. First strand cDNA from sexual stages was synthesized with a primer that specifically hybridizes to either anti-sense message (A) for *pgs28* (lane 1,2,5,6,9,and 10); or a primer that binds to sense message (S) for *pgs28* (lane 3,4,7,8,11,12). cDNA was synthesized from total RNA obtained from purified gametes and zygotes immediately (lane 1-4), 24 hours (lane 5-8) and 48 hours (lane 9-12) after isolation. RT-PCR was performed on all samples using the same primer pair. RT minus samples were also PCR amplified as negative controls (lanes 2,4,6, 8,10, 12). pBR322 MspI digest (lanes M).


Search for 14bp tag on NCBI database and compiled database of *P.falciparum* genomic sequence
Multiple matches
18% (35/187)
Single Match
70% (130/187)
No match
12% (22/187)
Un-annotated sequence
68% (88/130)
Annotated genes
32% (42/130)
Search for 13bp tag on NCBI and compiled databases
BLASTx against entire NCBI database
Unknown ORFs and Hypothetical proteins
(39/88)
Known ORFs (49/88)
-Predicted proteins
-Pf homologues of known proteins
Single match
(4/22)
No match
(18/22)



Percentage frequency	Total number of tags	Total number of genes	Matches to *P.falciparum* GenBank database
>1.1	96	1	1 (100%)
0.2-0.6	441	14	10 (71%)
0.05-0.2	1307	172	98 (57%)
0.025-0.05	2696	1092	ND
<0.025	3795	3587	ND
Total	8335	4866	ND

Fig. 3

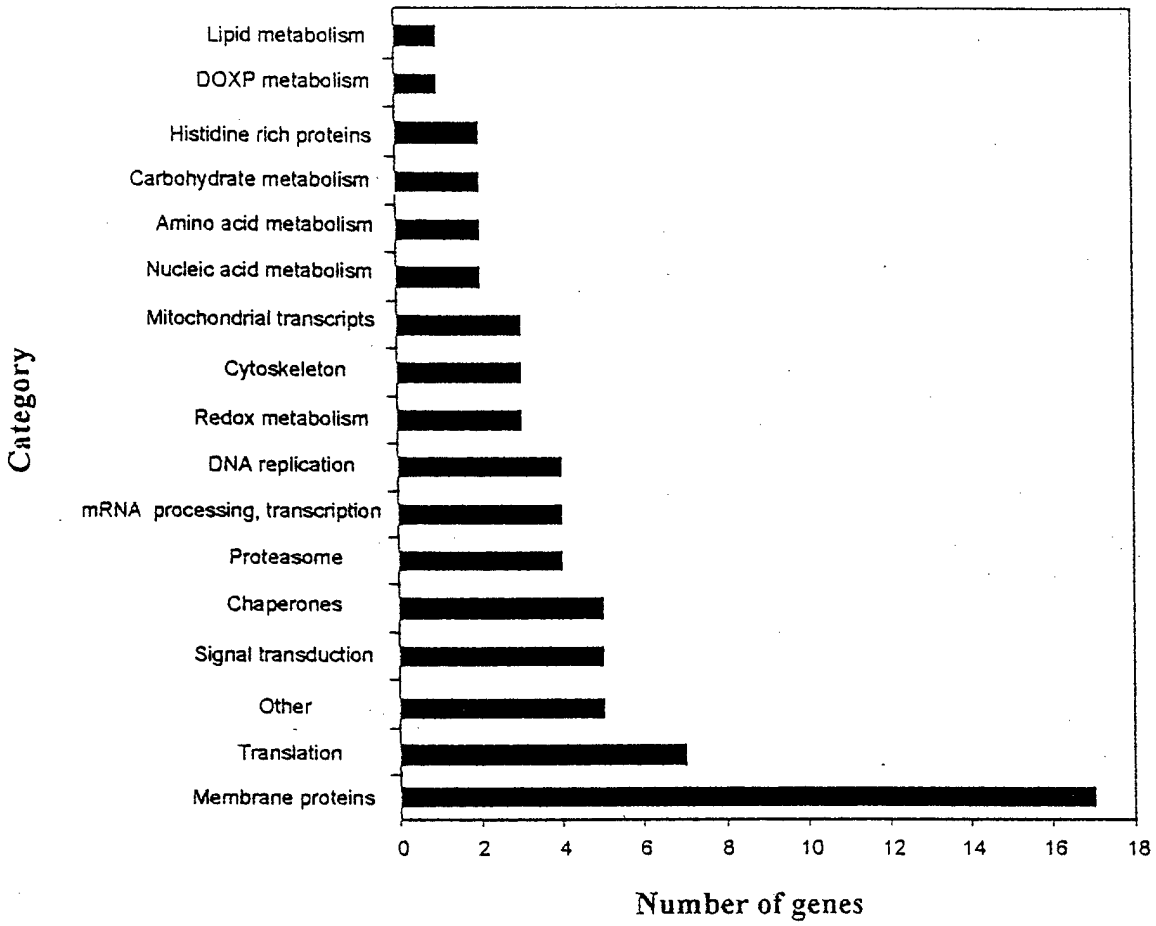
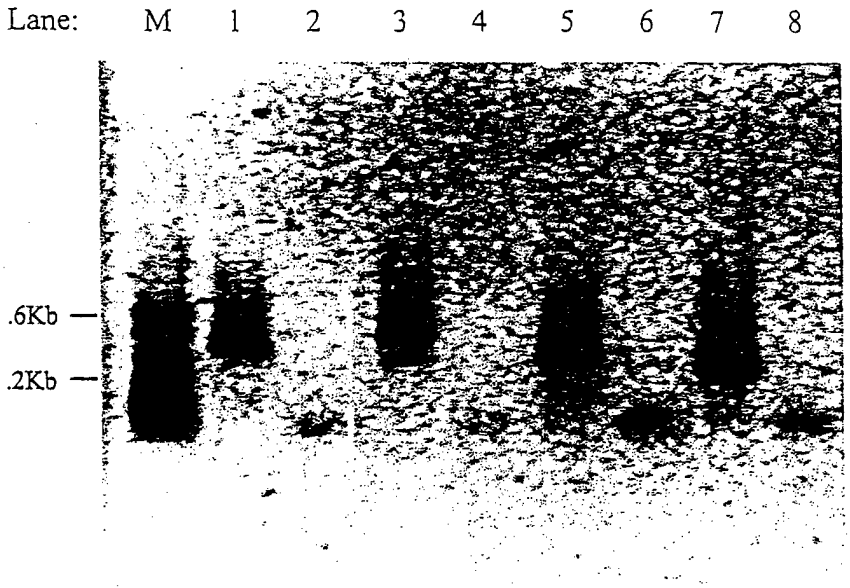
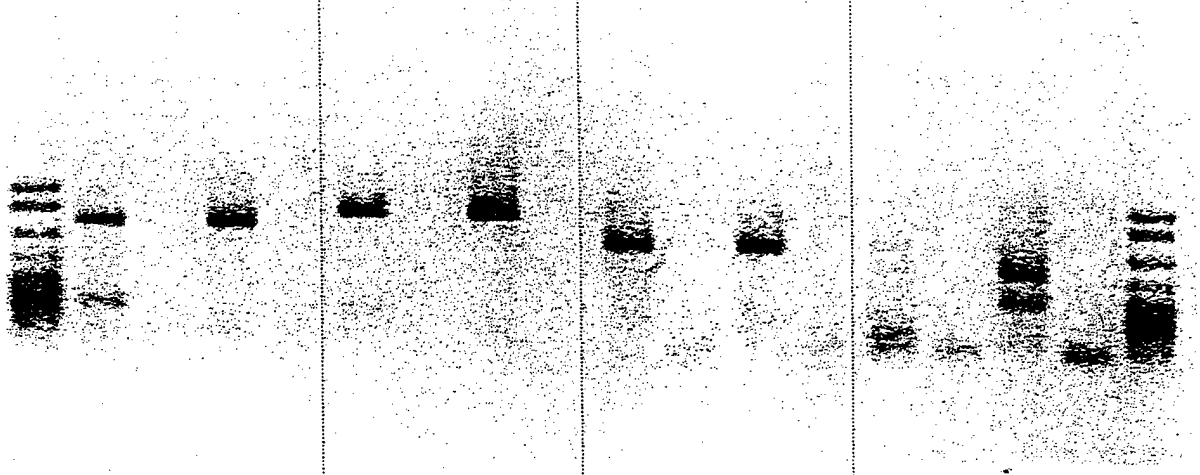

Lipid metabolism
DOXP metabolism
Histidine rich proteins
Carbohydrate metabolism
Amino acid metabolism
Nucleic acid metabolism
Mitochondrial transcripts
Cytoskeleton
Redox metabolism
DNA replication
mRNA processing, transcription
Proteasome
Chaperones
Signal transduction
Other
Translation
Membrane proteins
Category
0
2
4
6
8
10
12
14
16
18
Number of genes

Fig 4



Gene:	pfg27/25	rap-1	msp-3	calmodulin
Northern analysis:	+	+	+	+
SAGE tag counts:	20	13	15	6
Quantitative Northern: (femtomoles of transcript)	ND	ND	6	2

gene:		*pfg27/25*				*rap-1*				*msp-3*				calmodulin				
med cDNA strand:		A	A	S	S	A	A	S	S	A	A	S	S	A	A	S	S	
Lane:	M	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	M



Gene identity confirmed by sequence:	-	+	+	+	+	+	+	+
Anti-sense predicted by SAGE:	no		yes		yes	yes		

RNA:	S	A	3D7	S	A	3D7	S	A	3D7	S	A	3D7
Probe:	S	S	S	A	A	A	S	S	S	A	A	A
Lanes:	1	2	3	4	5	6	7	8	9	10	11	12

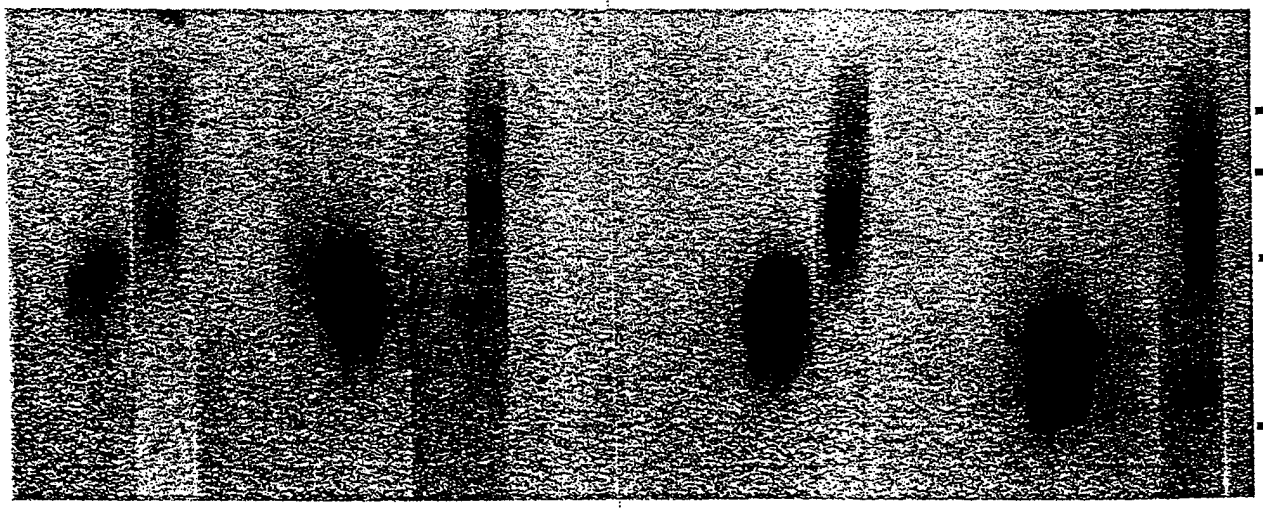


Fig 6

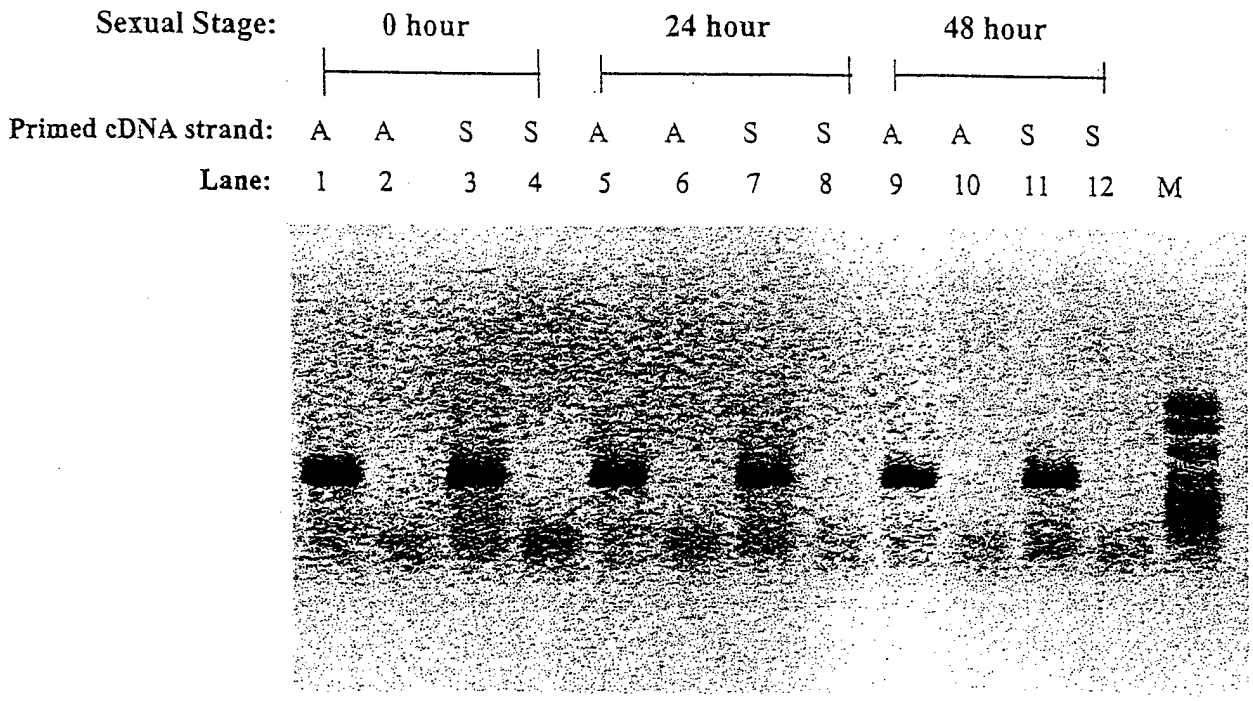

Sexual Stage:
0 hour
24 hour
48 hour
Primed cDNA strand:
A A S S A A S S A A S S
Lane:
1 2 3 4 5 6 7 8 9 10 11 12 M

Table I. Highly expressed genes in the 3D7 library

Tag	% Abundance	Gene description
TCAGGCGTTA	1.18	mitochondrial 6Kb product
GTGGTGGTGC	0.70	no match to database
GAGCAAGCAG	0.58	unknown protein
GAAGTCGAAA	0.44	5.8S ribosomal RNA
ATTTGAAGCA	0.36	Rhop H3
CTAAAGCACC	0.28	ras -related nuclear protein
TTGAAGCTGA	0.26	heat shock protein-70
AACGACAAGA	0.25	Pfg27/25
CCAAATGATG	0.25	polyubiquitin
TACAGCTGCT	0.18	merozoite surface protein-3
GGAAATAAAG	0.17	tartarate resistant acid phosphatase
TGAGTCAAAC	0.17	no match to database
GGCACAACTA	0.17	thioredoxin
TAAACTTTTG	0.16	rhoptry-associated protein-1
TTGTTTCATA	0.09	rifin
CGAGTAAAAG	0.09	1-deoxy-D-xylulose 5-phosphate synthase
CCAACTAAGG	0.07	ATPase subunit B
TGATGGCTTG	0.07	ornithine decarboxylase
TTCCGAACTT	0.07	triose phosphate isomerase
AGAGATCCGC	0.06	ubiquinol cytochrome-c reductase like protein
GATACTCTTG	0.06	26S proteasome beta subunit

Tag represents the 10bp SAGE tag sequence adjacent to the NlaIII site.
Abundance is listed as a percentage of all 8335 tags in the SAGE library